Etude de l'influence du passage à l'échelle sur les modèles de recherche d'information

Amélie Imafouo

Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

N° d'ordre : 421 I

THESE

Présentée par

Amélie IMAFOUO

pour obtenir le grade de
Docteur de l'Ecole Nationale Supérieure des Mines de Saint-Etienne

spécialité informatique

Etude de l'influence du passage à l'échelle sur les modèles de recherche d'information

Soutenue à Saint-Etienne le 6 décembre 2006
En présence d'un jury composé de :

MARIE FRANCINE MOENS	Professeur, Université Catholique de Louvain(Belgique) Examinatrice et Présidente
MOHAND BOUGHANEM	Professeur, Institut de Recherche en Informatique (Toulouse) Rapporteur
ERIC GAUSSIER	Professeur, Université Joseph Fourrier (Grenoble) Rapporteur
PATRICE BELLOT	Maître de Conférence, Université d'Avignon et des Pays de Vaucluse Examinateur
MICHEL BEIGBEDER	Maître-Assistant, Ecole des Mines de Saint-Etienne Directeur de recherche de thèse
JEAN-JACQUES GIRARDOT	Maître de Recherche, Ecole des Mines de Saint-Etienne Directeur de thèse
XAVIER BAY	Maître-Assistant, Ecole des Mines de Saint-Etienne invité

● **Spécialités doctorales :** **Responsables :**

SCIENCES ET GENIE DES MATERIAUX	**J. DRIVER** Directeur de recherche – Centre SMS
MECANIQUE ET INGENIERIE	**A. VAUTRIN** Professeur – Centre SMS
GENIE DES PROCEDES	**G. THOMAS** Professeur – Centre SPIN
SCIENCES DE LA TERRE	**B. GUY** Maitre de recherche
SCIENCES ET GENIE DE L'ENVIRONNEMENT	**J. BOURGOIS** Professeur – Centre SITE
MATHEMATIQUES APPLIQUEES	**E. TOUBOUL** Ingénieur
INFORMATIQUE	**O. BOISSIER** Professeur – Centre G2I
IMAGE, VISION, SIGNAL	**JC. PINOLI** Professeur – Centre CIS
GENIE INDUSTRIEL	**P. BURLAT** Professeur – Centre G2I
MICROELECTRONIQUE	**Ph. COLLOT** Professeur – Centre CMP

● **Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat** (titulaires d'un doctorat d'Etat ou d'une HDR)

BENABEN	Patrick	PR 2	Sciences & Génie des Matériaux	SMS
BERNACHE-ASSOLANT	Didier	PR 1	Génie des Procédés	CIS
BIGOT	Jean-Pierre	MR	Génie des Procédés	SPIN
BILAL	Essaïd	MR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR 2	Informatique	G2I
BOUDAREL	Marie-Reine	MA	Sciences de l'inform. & com.	DF
BOURGOIS	Jacques	PR 1	Sciences & Génie de l'Environnement	SITE
BRODHAG	Christian	MR	Sciences & Génie de l'Environnement	SITE
BURLAT	Patrick	PR 2	Génie industriel	G2I
COLLOT	Philippe	PR 1	Microélectronique	CMP
COURNIL	Michel	PR 1	Génie des Procédés	SPIN
DAUZERE-PERES	Stéphane	PR 1	Génie industriel	CMP
DARRIEULAT	Michel	ICM	Sciences & Génie des Matériaux	SMS
DECHOMETS	Roland	PR 2	Sciences & Génie de l'Environnement	SITE
DELAFOSSE	David	PR 2	Sciences & Génie des Matériaux	SMS
DOLGUI	Alexandre	PR 1	Informatique	G2I
DRAPIER	Sylvain	PR 2	Mécanique & Ingénierie	CIS
DRIVER	Julian	DR	Sciences & Génie des Matériaux	SMS
FOREST	Bernard	PR 1	Sciences & Génie des Matériaux	SMS
FORMISYN	Pascal	PR 1	Sciences & Génie de l'Environnement	SITE
FORTUNIER	Roland	PR 1	Sciences & Génie des Matériaux	CMP
FRACZKIEWICZ	Anna	MR	Sciences & Génie des Matériaux	SMS
GARCIA	Daniel	CR	Génie des Procédés	SPIN
GIRARDOT	Jean-Jacques	MR	Informatique	G2I
GOEURIOT	Dominique	MR	Sciences & Génie des Matériaux	SMS
GOEURIOT	Patrice	MR	Sciences & Génie des Matériaux	SMS
GRAILLOT	Didier	DR	Sciences & Génie de l'Environnement	SITE
GROSSEAU	Philippe	MR	Génie des Procédés	SPIN
GRUY	Frédéric	MR	Génie des Procédés	SPIN
GUILHOT	Bernard	DR	Génie des Procédés	CIS
GUY	Bernard	MR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HERRI	Jean-Michel	PR 2	Génie des Procédés	SPIN
KLÖCKER	Helmut	CR	Sciences & Génie des Matériaux	SMS
LAFOREST	Valérie	CR	Sciences & Génie de l'Environnement	SITE
LE COZE	Jean	PR 1	Sciences & Génie des Matériaux	SMS
LI	Jean-Michel	EC (CCI MP)	Microélectronique	CMP
LONDICHE	Henry	MR	Sciences & Génie de l'Environnement	SITE
MOLIMARD	Jérôme	MA	Sciences & Génie des Matériaux	SMS
MONTHEILLET	Frank	DR 1 CNRS	Sciences & Génie des Matériaux	SMS
PERIER-CAMBY	Laurent	MA1	Génie des Procédés	SPIN
PIJOLAT	Christophe	PR 1	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR 1	Génie des Procédés	SPIN
PINOLI	Jean-Charles	PR 1	Image, Vision, Signal	CIS
STOLARZ	Jacques	CR	Sciences & Génie des Matériaux	SMS
SZAFNICKI	Konrad	CR	Sciences de la Terre	SITE
THOMAS	Gérard	PR 1	Génie des Procédés	SPIN
TRAN MINH	Cahn	MR	Génie des Procédés	SPIN
VALDIVIESO	Françoise	CR	Génie des Procédés	SPIN
VALDIVIESO	François	MA	Sciences & Génie des Matériaux	SMS
VAUTRIN	Alain	PR 1	Mécanique & Ingénierie	SMS
VIRICELLE	Jean-Paul	CR	Génie des procédés	SPIN
WOLSKI	Krzysztof	CR	Sciences & Génie des Matériaux	SMS
XIE	Xiaolan	PR 1	Génie industriel	CIS

Glossaire :

PR 1	Professeur 1ère catégorie
PR 2	Professeur 2ème catégorie
MA(MDC)	Maître assistant
DR 1	Directeur de recherche
Ing.	Ingénieur
MR(DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
ICM	Ingénieur en chef des mines

Centres :

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
SITE	Sciences Information et Technologies pour l'Environnement
G2I	Génie Industriel et Informatique
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

Remerciements

Je voudrais tout d'abord remercier le Ciel pour les grâces dont Il m'a comblé

Ma profonde gratitude va à l'endroit de mes deux rapporteurs de thèse, le Pr Mohand BOUGHANEM et le Pr Eric GAUSSIER, pour leurs remarques très pertinentes et leurs nombreuses suggestions utiles.

Un grand merci au Dr Patrice BELLOT qui a accepté d'être examinateur de mon travail et membre de mon jury de thèse pour ses commentaires constructifs.

Merci particulièrement au Pr Marie-Francine MOENS pour s'être déplacée de Belgique dans le cadre de ma soutenance et pour avoir examiné ma thèse et présidé mon jury.

Un grand merci au Dr Xavier BAY pour son aide pour la prise en compte de techniques statistiques dans mon travail et pour sa grande disponibilité.

Je remercie grandement le Dr Michel BEIGBEDER, directeur de recherche de ce travail, pour la confiance qu'il m'a accordée au cours de ces années de thèse.

Merci au Pr Jean-Jacques GIRARDOT, directeur de ma thèse pour m'avoir permis d'effectuer ce travail au sein de son équipe

Merci aussi à toutes les rencontres enrichissantes faites au sein de l'équipe RIM et du centre G2I, au contact du personnel et de tous mes ami(e)s et collègues doctorants. Merci pour les bons moments passés ensemble, particulièrement à Annabelle et Olga.

Je remercie grandement tous ceux et celles, bien nombreux pour que ces pages me permettent de les citer tous, qui m'ont soutenu et encouragé de diverses façons.

Ma'a Moni et Pa'a Jean, soyez honorés pour tous les sacrifices consentis pour mon éducation

Dr Jean-Claude HOCHON, ton amitié m'honore et elle est un plaisir. La pertinence de tes conseils sur tous les plans (et de tes blagues) m'a toujours soutenu et guidé dans mon initiation à la recherche.

Dr Edwige BADJIO, merci de tes encouragements permanents depuis mon arrivée en France, au délà de tes difficultés personnelles.

Merci à toi Colince NGATSI : ta bonne humeur m'accompagne au quotidien, ta confiance et ton affection me fortifient.

Merci à Cendrine et Hervé MOGTO pour leur accueil à St-Etienne et leur amitié très édifiante

Hondjack DEHAINSALA, William MOCTO, Alain NKENGNE, Thierry NITCHEU, Félicien TALLA, François KOUAKOU, merci pour tant de choses : les apéros, les repas, les heures au téléphone, tous les bons moments qui m'ont aidé à me revigorer pour ma thèse.

En signe d'encouragement à toute personne que la quête du savoir amène à vivre loin des siens. A mes frères et sœurs : Puisse cette contribution à la recherche scientifique augmenter en vous la détermination dans votre quête de la connaissance.

Table des matières

Chapitre 1

Introduction

Sommaire

Les systèmes d'information informatisés couvrent progressivement toute la production d'information, soit directement lors de la production (cas des dessins animés en images de synthèse) ou après numérisation. Dans le même sens, nul ne peut nier les changements induits par l'explosion de l'utilisation des réseaux d'échange et de communication et notamment de l'Internet. Cette explosion engendre une croissance très significative des applications destinées à permettre à tout utilisateur de devenir producteur d'information et en même temps de pouvoir accéder aisément à cette information. En intégrant, de plus, les effets de la mise en œuvre d'intranets et d'extranets d'entreprises et d'organisations diverses ainsi que les bibliothèques numériques, on prend conscience du volume sans cesse grandissant d'information disponible sous forme numérique et du nombre croissant d'utilisateurs auxquels est confronté tout système d'information de

nos jours.

En effet, de plus en plus d'administrations, d'associations, d'entreprises, d'universités et d'écoles s'appuient sur l'utilisation de l'information numérique. Les développements technologiques rapides dans les domaines de la transmission, du stockage et de la sécurité des données engendrent le besoin (et la possibilité) de créer toujours plus d'informations (ou de générer des nouvelles informations à partir d'informations existantes). Ainsi, le volume d'informations ne se mesure plus en giga-octets sur la Toile par exemple mais en téra-octets voire en péta-octets et exa-octets[1]*. Parmi les disciplines dont le cœur de métier porte sur le stockage et la représentation de l'information, se trouve la recherche d'information. Elle a comme objectif majeur de fournir des moyens d'accès aisé à la masse d'information sans cesse en augmentation. D'autres disciplines visent un ou plusieurs des objectifs de la RI tout en se distinguant d'elle à certains points de vue. Nous commençons donc par cadrer ce que nous entendons par recherche d'information (et ce qui n'en est pas selon nous) en introduisant ses concepts clés, concepts dont nous débattrons au long des chapitres de cette thèse. Nous posons ensuite la problématique de la Recherche d'Information face à la croissance en quantité de l'information numérique, qui est l'objectif des nos travaux de recherche présentés ici. Nous terminons cette introduction générale par un tour d'horizon de nos contributions et par la description du contenu des prochains chapitres.*

1.1 La Recherche d'Information

D'après le Dictionnaire Larousse, le mot Information désigne un *élément de connaissance susceptible d'être codé pour être conservé, traité ou communiqué.*

Les bibliothécaires et les documentalistes ont longtemps eu pour rôle de permettre aux usagers d'accéder facilement à un ensemble de livres et d'écrits divers. Ces derniers mettaient alors en place des catalogues de l'ensemble des livres accessibles, en les listant par exemple suivant un ordre alphabétique sur les titres ou en les regroupant par auteur ou par date de parution, ce qui permettait un accès facilité aux livres. Avec l'avènement de l'outil informatique, il a été possible d'automatiser ces techniques et ceci a donné lieu à des domaines comme les Bases de Données, les Bibliothèques Numériques, la Recherche d'Information que nous abrégerons souvent RI. Ce dernier domaine est celui qui nous intéresse. Parmi les premiers travaux de ce domaine, on peut citer ceux de *Mooers* en 1960 qui définit la RI comme *le processus ou la méthode par lequel un utilisateur prospectif d'information est capable de convertir son besoin d'information en une liste*

[1] Ces mesures sont des puissance de deux : giga=$2^{30}(\simeq 10^9)$, tera=$2^{40}(\simeq 10^{12})$, puis péta, exa, zetta et yotta=$2^{80}(\simeq 10^{24})$

réelle de citations de documents accessibles et contenant de l'information utile pour lui[2] ([111], page 25).

Du fait qu'elle porte sur des documents, la RI est à distinguer des autres domaines comme les bases de données qui stockent, traitent et restituent aussi de l'information. La RI a connu un essor particulier avec l'avènement de l'Internet qui a apporté avec lui une croissance fulgurante des informations numériques. Pour en faciliter l'accès, les systèmes de RI constituent aujourd'hui le moyen le plus utilisé. Nous présentons les concepts de base de la RI, les phases d'un processus traditionnel de RI et nous mettons en exergue des disciplines qui chevauchent la RI en étant toutefois distinctes d'elle.

1.1.1 Concepts de base de la Recherche d'Information

Document : Emprunté du latin documentum, exemple, modèle, enseignement, ce qui sert à instruire. Acte écrit qui sert de témoignage, de preuve.

Requête : Expression formalisée d'une demande.

Pertinence : Qualité de ce qui est juste, judicieux, approprié.

Dans le cas général, pour parler de Recherche d'Information, au moins deux pré-requis doivent être vérifiés :

- le premier pré-requis est l'existence d'un besoin d'information
- le deuxième pré-requis est de disposer d'un ensemble de documents (contenant potentiellement l'information souhaitée) au sein duquel va s'opérer la recherche.

Une fois ces deux contraintes réunies, il faut adopter une stratégie pour rechercher l'information qui est attendue au sein de l'ensemble des documents. Il est donc nécessaire de formuler le besoin d'information sous une forme compatible à la stratégie de recherche adoptée ; cette dernière devant fournir la liste des informations qu'elle a estimées intéressantes.

En recherche d'information, trois aspects sont donc incontournables :

- tout d'abord, le besoin d'information : il est communément exprimé sous la forme d'une *requête*, la façon de l'exprimer est appelée *langage de requête* ;
- ensuite, l'ensemble des documents : c'est l'espace de recherche aussi appelée collection ou corpus ;
- enfin, la stratégie de recherche : elle est sous-tendue par un *modèle de recherche d'information.*

[2]Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored information.

Les informations contenues dans l'espace de recherche sont couramment organisées sous formes d'entités nommées *documents*. Ce sont des entités qui encapsulent l'information. Elles sont minimales dans le sens où ce sont les résultats retournés par le système.
Le résultat retourné au terme de la recherche est donc aussi généralement nommé document. Toutefois, dans le cadre de certaines applications, l'unité d'information retournée comme résultat peut être une partie d'un document (cas de la recherche de passages, cas de la recherche dans les documents structurés XML).

La recherche d'information a pour objectif la définition de modèles et de systèmes dédiés à la représentation des documents, permettant l'accès à ces documents en vue de satisfaire les besoins d'information des usagers. Cette définition introduit le concept d'*usager* ou *utilisateur*, qui représente l'entité humaine qui est en quête d'une information. La place de l'usager est centrale puisque c'est lui qui formule/crée le besoin d'information et c'est lui qui détermine si les résultats retournés lui apportent satisfaction ou non. Dans nos travaux, nous nous intéressons exclusivement à la RI automatisée par utilisation des ordinateurs. Les documents manipulés sont donc numériques et les outils qui permettent la mise en œuvre de la stratégie de recherche sont des outils informatiques.
Ainsi un document est une entité atomique qui peut être recherchée, retrouvée et consultée, sans pour autant être nécessairement physiquement sauvegardée comme une entité unique indépendante. C'est donc ce qui véhicule l'information qui peut prendre plusieurs formes (texte, image, son, vidéo, etc). Concrètement, un document pourra donc être un article de journal, une section ou un chapitre d'un livre, une phrase, etc.
Une requête est le véhicule du besoin d'information de l'usager. En général en RI, la requête est un ensemble de mots-clés, ordonnés ou non, pondérés ou non, liés éventuellement par des opérateurs par exemple booléens. Différentes classes de requêtes différentes ont été déterminées (mots-clés uniquement, requêtes dans des sous-parties de documents, etc.). Ce sont les langages de requêtes qui définissent la forme des requêtes (requêtes booléennes, mots clés pondérés, requêtes structurées (langage XQuery, XPath), requêtes sous forme d'expressions régulières, langage naturel etc.).
Ces différents concepts sont manipulés durant le processus de RI qui se divise en plusieurs phases décrites dans la section suivante.

1.1.2 Processus de Recherche d'Information

Première phase : la création/collecte du corpus Comme évoqué précédemment, un des pré-requis pour démarrer un processus de RI est de disposer d'un ensemble d'informations au

sein desquelles va s'opérer la recherche. Naturellement donc, la première phase de tout processus de RI est la mise en place de la collection de documents constituant l'espace de recherche. Pour des systèmes de RI sur le Web (couramment appelés moteurs de recherche), la collecte des informations se fait de façon périodique au travers des *crawlers* qui parcourent (et aspirent) les pages des sites web en suivant les liens hyper-textes. Dans le monde académique, pour permettre des études reproductibles sur des ensembles de documents proches du monde réel, des collections (dites de test) sont construites depuis plusieurs années, certaines sont constituées d'articles de journaux, d'autres sont des extraits du Web, etc. Ces collections sont utilisées pour l'évaluation et la comparaison des systèmes de RI, dans le cadre de campagnes d'évaluation comme *TREC* (Text REtrieval Conference [1]), *CLEF* (Cross Language Evaluation Forum [118]), *NTCIR* (NACCSIS-NII Test Collection for Information Retrieval [8]), etc.

Seconde phase : l'indexation Pour mener de façon efficace le processus de RI, il s'est avéré nécessaire de mettre en place une phase préalable d'analyse des documents, qui est indépendante du besoin d'information. En effet, il faudrait effectuer une recherche directement dans la collection de documents à chaque nouvelle requête. Un certain nombre de traitements peuvent être faits en amont ; il s'agit essentiellement de créer une représentation des documents qui rende aisé et rapide leur accès lorsque survient une demande de l'utilisateur. Cette seconde phase est l'*indexation*. La représentation des documents est une description souvent plus succincte que le contenu lui même. L'utilisation des représentants plutôt que des documents eux-mêmes peut avoir un impact sur la qualité de la recherche (perte d'information).
Pour construire la représentation des documents, différentes stratégies existent : la plus commune consiste à sélectionner les mots qui représentent le mieux le contenu sémantique du document et leur attribuer un poids indiquant leur importance dans le document. Cette stratégie peut être complétée par la suppression des mots *vides* (mots fréquents estimés sans grande portée sémantique, les pronoms, les conjonctions, prépositions, etc.), la *lemmatisation* (remplacer les mots par leur racine), la reconnaissance de termes complexes (utilisation de thésaurus, de dictionnaire,etc.). Ces traitements donnent naissance à un index qui est en général une structure permettant d'associer à chaque mot (terme) d'indexation la liste des documents le contenant, en plus éventuellement des informations comme le poids du terme, les positions de ses occurrences, les termes avec lesquels ils co-occurrent dans un document, etc.

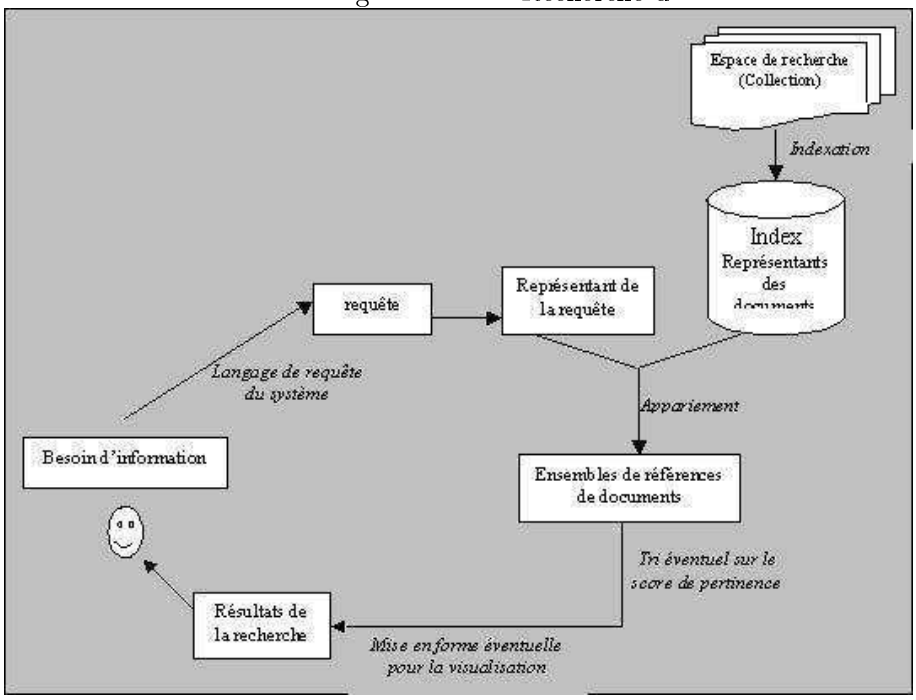
Troisième phase : l'interrogation Une fois l'index construit, il est possible de traiter les requêtes des utilisateurs. C'est la phase d'*interrogation* qui se déclenche à chaque requête. Cette phase nécessite tout d'abord la formulation du besoin d'information et ensuite la recherche de

l'information souhaitée dans l'ensemble des représentants des documents. Divers langages de requêtes existent mais dans tous les cas, la requête est toujours transformée en son représentant (comme les documents). Le langage doit permettre l'appariement entre la représentation des documents et celle de la requête. La fonction d'appariement détecte les documents *pertinents* pour la requête et les ordonne éventuellement suivant un *score de pertinence* qu'elle attribue à chaque document. Cette pertinence est celle du point de vue système et peut différer de celle du point de vue de l'utilisateur. Cette dernière peut avoir de nombreux sens et nous en discuterons dans le chapitre 3. La méthode de création de l'index et la fonction d'appariement sont les clés de voûte de tout *modèle de RI*. Les modèles traditionnels de RI sont présentés dans le second chapitre.

Quatrième phase : la visualisation *Le retour ou la visualisation des résultats* est une phase du processus souvent considérée comme partie de la phase d'interrogation. Elle consiste à rendre sous une forme intelligible par l'utilisateur, l'ensemble des documents que le système a estimé intéressants pour son besoin d'information. En général, une liste de références vers ces documents est fournie (lien hyper-texte pour les moteurs de RI sur le Web, numéro de document pour les systèmes de RI dans les campagnes d'évaluation). Il existe de plus en plus d'applications qui fournissent d'autres informations en plus des documents eux-mêmes (carte conceptuelle montrant les liens sémantiques entre les documents, estimation du taux de pertinence des documents, regroupement de documents par sites d'origine, etc.). Des extensions comme le retour de pertinence (*relevance feedback*), la reformulation de la requête peuvent enrichir cette phase.

Cinquième phase : l'évaluation Pour *mesurer* la qualité des résultats rendus par le système de RI, il est nécessaire de réaliser une *évaluation*. L'*évaluation* est une notion utilisée dans la plupart des domaines liés aux systèmes d'information. Elle consiste à établir une sorte de bilan permettant de déterminer si les objectifs initiaux du système ont été atteints. La RI n'échappe pas à ce principe. La phase d'évaluation en RI a fait l'objet de nombreuses études. Des métriques d'évaluation ont été mises en œuvre, et portent sur l'efficacité (les plus courantes étant la *précision* et le *rappel*) ou sur l'efficience (rapidité, ressources utilisées, temps de réponse). Les métriques liées à l'efficacité visent en général à établir le taux de documents pertinents ayant effectivement été retrouvés par le système de RI. Dans le monde académique qui utilise des collections statiques, il est nécessaire d'établir des jugements de pertinence (cette tâche est très souvent effectuée par des humains) sur les documents présents dans l'espace de recherche. Ceci permet d'avoir une base commune de jugements de pertinence pour évaluer les systèmes de RI. Les méthodes proposées pour faciliter l'évaluation et la comparaison de systèmes de RI sont abordées dans le second

FIG. 1.1 – Processus général de la Recherche d'Information



chapitre, ainsi que les problèmes que ces deux tâches soulèvent.

La figure 1.1 donne un aperçu des différentes phases de base d'un processus de RI. Chacune de ces phases est confrontée à la croissance exponentielle du volume d'information numérique. La prise en compte de cette croissance que nous appelons *passage à l'échelle en RI* est le thème de nos travaux.

1.2 Problématique et enjeux

1.2.1 La croissance d'information

La croissance exponentielle de l'information est liée à des causes diverses ; elle entraîne de nouveaux problèmes comme le risque de surabondance d'information qui peut entraîner à son tour la désorientation et la difficulté à trouver ce que l'on cherche. Les causes principales qui entraînent l'augmentation très rapide de l'information numérique sont liées à la nature même de cette information. Sa forme numérique permet son traitement par des ordinateurs de façon automatique. Les évolutions technologiques rapides en informatique (espace de stockage de plus en plus volumineux sur des supports de plus en plus miniaturisés, utilisation des réseaux, etc.) entraînent donc une production et une reproduction plus facile de cette information ainsi qu'une génération de nouvelles informations par le traitement d'informations existantes. La convivialité

des interfaces utilisateurs et la possibilité de créer, d'échanger de l'information rapidement et efficacement attirent également de plus en plus d'utilisateurs. L'analyse de ces causes et des chiffres tirés d'études quantitatives qui caractérisent cette croissance est faite dans le second chapitre.

1.2.2 Problématique

Dans un contexte industriel, les organisations et les entreprises accumulent des masses d'informations. Pour celles qui savent en tirer profit, ces masses d'informations sont un gisement. On voit en effet émerger des domaines comme la gestion des connaissances (*knowledge management*) alors que de nouvelles stratégies perpétuent des disciplines comme la fouille et l'analyse de données ou la gestion des bases de données. Tous ces domaines ont en commun dans leur cœur de métier la manipulation de l'information existante pour en tirer de nouvelles informations (connaissances) et permettre des prises de décision stratégiques. La discipline dédiée particulièrement à faciliter l'accès à cette masse d'information est la RI. Il est donc intéressant de se demander comment se comportent les outils/modèles mis en œuvre au sein de cette discipline face à la masse d'information qui augmente très vite. En effet, ce volume d'information comme tout gisement nécessite des stratégies adaptées afin d'exploiter au mieux son contenu. Il existe en effet des risques de surabondance, de dispersion, d'égarement par infopollution.

Notre travail se situe dans cette problématique de prise en compte du passage à l'échelle en RI. Nous commençons par un rapide tour de cette problématique par rapport aux cinq phases du processus de RI défini en section 1.1.2, avant de donner les phases sur lesquelles notre travail a porté. Pour ce qui est des phases de mise en place de l'espace de recherche et de création d'index, il est nécessaire de réduire au mieux les espaces de stockage utilisés, tout en s'assurant que cela ne crée pas des freins dans les autres phases comme l'interrogation. Les travaux existants se sont surtout intéressés à ces deux phases en fournissant des moyens de compression (logique ou physique) des données. En terme d'efficience, il existe également des études visant à réduire le temps d'indexation, le temps de mise en place des espaces de recherche sur le Web.
Pour ce qui est de l'interrogation, le temps de réponse est la clé du succès, couplé bien sûr à la pertinence des résultats. Ce temps de réponse dépend directement de la structure de représentation des documents dans l'index et de la *facilité* à l'apparier à la représentation de la requête. Il existe plusieurs types de fonction d'appariement correspondant aux différents modèles de RI existants. L'efficacité de chacune d'elles face à la croissance des collections mérite donc d'être étudiée, afin de mieux connaître leurs comportements et d'en déduire un ajustement possible de leurs paramètres pour une meilleure prise en compte du passage à l'échelle. Par exemple,

l'hétérogénéité est plus forte dans les grandes collections et ceci pourrait affecter l'intérêt des paramètres couramment utilisés dans les modèles de RI comme la fréquence documentaire. Il est à noter également pour l'interrogation que la prise en compte de l'utilisateur se fait aussi de plus en plus (RI interactive, retour de pertinence) et permettra de mieux appréhender son besoin d'information (qui est souvent peu clair pour lui même) pour mieux le satisfaire.
Cette satisfaction passe également par les techniques mises en œuvre pour la visualisation des résultats. Dans le domaine de l'Interaction Homme-Machine (IHM), de nombreux travaux portent sur les méthodes de visualisation de grandes quantités de données. Peu de ces méthodes ont été appliquées au domaine de la RI et les applications existantes manquent souvent d'évaluation. La prise en compte du passage à l'échelle à l'étape de retour des résultats au travers d'interfaces de visualisation adaptées est donc un axe qui pourra aider à mieux satisfaire l'utilisateur.
L'évaluation en RI ne tient pas toujours compte de la taille des informations traitées ; or il existe des modèles de RI pour lesquels la taille de collection est un paramètre de la fonction d'appariement. La croissance de collections peut induire des biais sur des paramètres utilisés dans ces modèles de RI. Il est donc crucial d'analyser l'influence de l'accroissement des collections sur les modèles de RI et de proposer des techniques d'évaluation qui prennent en compte la taille des collections.
La prise en compte du passage à l'échelle est donc un champ vaste en RI, sur lequel porteront certainement de nombreux travaux dans un futur proche. Nous présentons dans la section suivante les points sur lesquels nous nous sommes particulièrement penchés dans ce travail de recherche.

1.3 Objectifs et Synthèse des contributions

1.3.1 Objectifs de la thèse : description des axes de travail

Toutes les phases du processus de RI méritent d'être revues pour tenir compte du passage à l'échelle. Par rapport au passage à l'échelle, de nombreux travaux se sont intéressés à l'efficience en RI, par contre peu ont abordé la question de l'efficacité.
Nos travaux ont donc pour objectif d'étudier cet aspect quelque peu laissé pour compte dans les études actuelles. Pour ce faire, dans un premier temps sont proposées des techniques de création de collections permettant d'étudier le comportement des systèmes de RI (en terme d'efficacité/pertinence des résultats retournés).
Dans un second temps, des métriques pour évaluer la capacité des systèmes de RI à passer à l'échelle sont proposées. Face à une demande de l'utilisateur, un document retourné peut apporter plus ou moins de satisfaction ; la pertinence comme concept binaire nous semble à ce

titre limitative ; les métriques que nous proposons s'appliquent particulièrement pour analyser le comportement d'un système de RI quand il travaille sur une collection donnée qui croît au fur et à mesure (en nombre de documents), dans le cas où la pertinence d'un document face à une requête peut s'évaluer sur une échelle à plusieurs valeurs (pertinence non binaire).

1.3.2 Mise en œuvre de collections pour l'étude du passage à l'échelle

Le premier point sur lequel portent nos travaux est de fournir des moyens pour connaître le comportement des systèmes de RI face au passage à l'échelle. Une approche évidente pour étudier ce comportement est de faire tourner les systèmes de RI sur des collections de tailles croissantes. Toutefois, il ne suffit pas que les tailles des collections soient croissantes pour que l'on puisse tirer des conclusions sur les performances d'un système. Si les collections sont constituées de façon trop *différente*, il sera difficile (voire impossible) de dissocier dans les résultats obtenus sur les performances du système la part due à l'influence de la taille de la collection et la part due aux (contenu des) collections elles-mêmes. Par exemple, sur une collection de taille $T1$ contenant 100 documents pertinents, un système pourra avoir de meilleures performances que sur une collection de taille $T2$, ($T2 > T1$) contenant 60 documents pertinents. Peut-on en déduire que les performances du système se détériorent quand la taille de collection augmente ? Il est nécessaire de connaître et de contrôler le biais induit par ce qui constitue les collections. Pour effectuer ce contrôle, deux techniques de construction de collections de taille croissante sont proposées. Dans les deux techniques, le point de départ est une collection volumineuse dont nous avons/déterminons les *caractéristiques* ; des sous-collections sont construites et un contrôle est fait en cours de construction sur les caractéristiques de chaque sous-collection, celles-ci doivent être *similaires* aux caractéristiques de la collection initiale. Cette contrainte affranchit du biais lié à ce que contient chaque sous-collection, et permet de focaliser uniquement sur la taille.

- la première technique est reproductible, elle passe par la mise en place d'une collection dite *uniforme* et son intérêt est également qu'il n'y a pas de contraintes sur la façon de la découper en sous-collections pour réaliser des études ;
- la seconde technique utilise les méthodes statistiques d'échantillonnage de grands ensembles de données, notamment l'échantillonnage aléatoire sans remise. Nous obtenons donc des *échantillons* de tailles croissantes de la collection initiale. Il est alors possible de construire ensuite des *intervalles de confiance* sur les performances des systèmes de RI pour chaque taille d'échantillons ou d'utiliser des méthodes statistiques de visualisation graphique réputées efficaces comme les *box-plots* ; ces techniques de visualisation permettent de déduire des conclusions statistiquement valides sur le comportement d'un système de RI quand la

taille de collection augmente.

Chacune de ces méthodologies a été mise en oeuvre de façon expérimentale sur des collections de test standards du domaine de la RI. Le cas d'étude choisi pour ces expérimentations concerne l'évaluation des performances des systèmes en RI en terme de pertinence des résultats retournés. Ainsi, les *caractéristiques* dont nous avons tenu compte lors de la construction des sous-collections sont la proportion de documents pertinents, puisque la pertinence est au centre de l'évaluation. Les comportements de six systèmes de RI ont été analysés face à la croissance en taille des collections, en utilisant les métriques classiques utilisées de la RI.

1.3.3 Métriques pour évaluer le passage à l'échelle dans des environnements à pertinence multivaluée

Dans les environnements larges et modernes de RI, il est souhaitable d'avoir des SRI qui retournent des documents en fonction de leur degré de pertinence : par exemple d'abord les documents très pertinents, ensuite les documents pertinents et enfin les documents faiblement pertinents. Ce point de vue a été introduit par des travaux antérieurs aux nôtres, et nous le partageons. L'évaluation en RI doit donc pouvoir récompenser (ou tout au moins reconnaître) les SRI qui retournent les documents ayant le plus haut degré de pertinence en tête des autres documents. Pour ce faire, il est nécessaire de prendre en compte les différents degrés de pertinence d'un document par rapport à un besoin d'information. Les degrés de pertinence ont été étudiés dans des travaux antérieurs, bien que tous les auteurs s'accordent sur le fait que la pertinence est un phénomène d'une grande complexité car il induit des processus cognitifs et subjectifs. Nous avons proposé des métriques pour analyser et évaluer le comportement des modèles de RI lorsqu'on utilise des ensembles de recherche de plus en plus grands, en tenant compte de la pertinence multivaluée (chapitre 5). Soient deux collections C_1 et C_2 de taille croissante tels que $C_1 \subseteq C_2$ et un SRI S. On souhaite analyser le comportement de S sur chacune des deux collections ; notamment, on voudrait déterminer si les performances de S s'améliorent, restent stables ou se détériorent lorsque la taille de la collection augmente. Pour ce faire, nos métriques analysent le comportement des systèmes en se basant sur les degrés de pertinence des documents contenus dans la liste des résultats retournés suite à l'interrogation de chaque collection. Ces métriques permettent de discuter des notions comme l'importance d'un niveau de pertinence ou le gain d'information réalisé entre deux niveaux de pertinence. Nous avons proposé des formalisations de la notion de gain d'information entre deux niveaux de pertinence et formalisé les contraintes (évidentes) liées à l'attribution d'une valeur (numérique pouvant permettre de faire des calculs) pour caractériser un niveau de pertinence. Le problème de cette attribution en lui

même reste un problème non résolu en RI.

1.4 Description des chapitres

Ce document descriptif est construit autour d'une introduction générale, de quatres chapitres et d'une conclusion. Voici l'aperçu des éléments à venir :

- Après cette introduction générale, le second chapitre porte sur l'analyse détaillée du phénomène de croissance d'information qui est l'essence même de ce travail de recherche. Les causes diverses qui l'engendrent et des études quantitatives qui mettent des chiffres sur ce phénomène sont analysés. Nous parcourons ensuite les différentes phases d'un processus classique de recherche d'information pour analyser l'influence du passage à l'échelle et commenter les travaux ayant fourni des pistes pour prendre en compte cette influence.
- le troisième chapitre traite de l'évaluation en recherche d'information. Le concept de pertinence qui est au centre de cette évaluation est présenté sous différentes facettes, tantôt considéré comme un concept binaire, tantôt comme un concept pouvant prendre plusieurs valeurs. Les techniques d'évaluation utilisées pour la RI (en milieu académique) et un tour d'horizon des métriques existantes, qu'elles soient à pertinence binaire ou multivaluée est également réalisé ;
- le quatrième chapitre est dédié à la présentation des techniques de création de sous-collections que nous proposons pour l'étude du passage à l'échelle. Ces techniques visent à partir d'une collection initiale pour construire des sous-collections de taille croissante mais ayant des *caractéristiques similaires* à celle de la collection initiale. Elles permettent de s'affranchir du biais du contenu des sous-collections pour se focaliser uniquement sur leur taille. Nous présentons ensuite le cas d'étude réalisé sur l'évaluation en RI et les expérimentations menées en utilisant les métriques classiques de la RI ;
- dans le cinquième chapitre, nous discutons des métriques que nous proposons pour l'évaluation du passage à l'échelle dans les environnements à pertinence multivaluée. Une comparaison entre ces métriques et des métriques précédemment développées en RI est également établie ;
- nous terminons ce rapport par une conclusion qui fait une synthèse des informations présentées dans les différents chapitres sus-cités. Après avoir résumé les apports de ce travail de recherche, les limites découvertes sont mentionnées et nous mettons un point final par les perspectives à court terme de nos travaux ainsi que les horizons futurs dans un plus long terme.

Chapitre 2

Le passage à l'échelle dans les phases du processus de RI

Sommaire

FIG. 2.1 – Différence entre une donnée et une information

Donnée	+ Modèle d'interprétation	= information
1 56 05 99 131 088	+ modèle des numéros de sécurité sociale en France	= *l'identifiant d'un homme né en* 1956 *à l'Etranger*

On qualifie d'information toute donnée que le système nerveux central est capable d'interpréter pour se construire une représentation du monde et pour interagir correctement avec lui. De nos jours, de nombreux facteurs facilitent une production croissante d'information. Nous nous intéressons dans ce chapitre à cette croissance d'information en analysant ses causes et présentons quelques chiffres significatifs qui la caractérisent dans différents domaines qui sont proches de la Recherche d'Information. Nous nous attardons ensuite sur les travaux de recherche qui se sont intéressés à la prise en compte de l'effet de cette masse grandissante d'information sur les phases d'un processus de RI que sont la création de l'espace de recherche, l'indexation des données contenues dans cet espace, l'interrogation de ces données grâce à des requêtes et la visualisation des résultats fournis en réponse par les systèmes de RI.

2.1 Introduction

Une donnée est une description élémentaire, souvent codée, d'une chose ou d'un événement. Elle peut être conservée et classée sous forme de texte alphanumérique, d'image, de son. Une donnée en elle même n'a pas de signification dans l'absolu. Il est nécessaire de lui associer un modèle d'interprétation qui permet de lui attribuer un sens (et une utilité).

On qualifie d'information toute donnée que le système nerveux central est capable d'interpréter pour se construire une représentation du monde et pour interagir correctement avec lui. *Dervin et Nilan* [40] cité par *Spink et al.* [147] parlent d'un phénomène subjectif construit par les humains dans un processus de construction de sens (*sense-making process*). L'information est ainsi constituée d'une donnée et d'un modèle d'interprétation de cette donnée. Ainsi la suite de chiffres 1 56 05 99 131 088 peut être interprétée dans le modèle des numéros de sécurité sociale en France comme *l'identifiant d'un homme né en Mai* 1956 *à l'étranger* (voir figure 2.1). Le dictionnaire Larousse définit le terme information comme *un élément de connaissance susceptible d'être codé pour être conservé, traité ou communiqué*, donnant ainsi les principales fonctions de tout système ayant en son cœur l'information. Pour une meilleure organisation, ces informations sont souvent

assemblées et stockées pour y avoir accès plus tard, organisées pour faciliter leur accès, traitées pour générer de nouvelles informations et restituées en cas de besoin.

Le développement technologique a apporté la notion de système d'information, qui peut être défini comme un système constitué du matériel, des procédures, des ressources humaines, ainsi que des données qui y sont traitées, et dont le but est de fournir de l'information. Ce système d'information est l'entité qui va désormais permettre l'accès aux informations. Le système d'informations informatique fait son apparition avec l'apparition de l'informatique. Il permet de réaliser toutes les manipulations sur les informations citées précédemment de façon plus ou moins automatique, au travers d'un ordinateur. L'informatique apporte également avec elle la notion d'information numérique. L'information numérique est tout particulièrement facile à stocker, à reproduire et à transmettre. L'automatisation du traitement rend également plus aisée la génération de nouvelles informations numériques à partir d'informations existantes. Ainsi, les rapides et nombreuses évolutions en informatique facilitent de plus en plus la production d'information numérique et sa mise à disposition au travers des réseaux d'échange et de communication. La croissance en volume de l'information stockée de façon numérique est donc quasi-exponentielle ces dernières années. Des études se sont intéressées de différentes façons à ce phénomène de croissance en analysant des raisons qui facilitent cette croissance ou en quantifiant la masse d'information produite. La première partie de ce chapitre s'intéresse à cette croissance d'information en parcourant ces travaux.

Tout ce volume d'information nécessite d'être stocké, traité et analysé mais il est surtout primordial de permettre un accès aisé à l'information particulière que recherche un utilisateur, d'où la mise en œuvre de multiples systèmes de RI. Chacun de ces outils de RI a pour but de faciliter l'accès des utilisateurs à une partie de l'information totale stockée qui est censée les intéresser particulièrement. Le processus de recherche de cette information se fait en plusieurs étapes. Tout d'abord, il faut collecter l'ensemble des informations qui vont constituer l'espace de recherche ; pour faciliter l'accès aux informations collectées, il est nécessaire de les indexer. Il faut ensuite prendre en compte le besoin d'information de l'utilisateur ; pour ce faire, le système de RI recueille ce besoin d'information, par exemple (cas le plus simple) sous la forme d'une suite de mots clés. Le système de recherche réalise ensuite l'appariement entre le besoin d'information de l'utilisateur et les éléments contenus dans l'index des informations collectées : les éléments qui sont les plus intéressants d'après le système de RI sont alors retournés à l'utilisateur. Ce retour de résultats doit se faire sous une forme accessible, compréhensible et exploitable pour l'utilisateur. Ces étapes basiques d'un processus de RI peuvent éventuellement être accompagnées d'autres étapes comme l'aide à la formulation du besoin d'information, le retour de pertinence, etc. La

croissance du volume d'information agit sur chacune des phases du processus de RI.
Nous nous intéressons donc dans les deuxième, troisième et quatrième sections de ce chapitre à l'impact de la croissance d'information sur les étapes d'un processus de recherche d'informations. La dernière section de ce chapitre donne une synthèse des éléments abordés et nous introduit au prochain chapitre.

2.2 La croissance d'information

Informations numériques : gisement exploitable ou amas de données ?

L'information numérique est au cœur de nombreux domaines tel que les bases de données, les bibliothèques numériques, le web, la RI. Elle est la principale richesse des organisations modernes, sous forme de technologies, de savoirs, de savoir-faire, de brevets, de compétences métier, de stratégies ou autres. La croissance exponentielle de cette information est liée à des causes diverses que nous analysons ; nous nous intéressons également à son impact sur les phases du processus de RI.

2.2.1 Causes

Les causes principales qui entraînent l'augmentation exponentielle de l'information numérique sont liées à sa nature numérique qui permet son traitement par des ordinateurs de façon automatique. La production de plus en plus aisée de l'information numérique, la possibilité de l'échanger rapidement et efficacement attirent également de plus en plus d'utilisateurs.

Evolutions technologiques (espaces de stockage, réseaux)

Les évolutions technologiques concernent tout d'abord les espaces de stockage d'information. Il y a une dizaine d'années encore, les disquettes étaient les principales unités de stockage externes et les disques durs n'excédaient pas quelques dizaines ou centaines de méga-octects. Le progrès a été rapide et de nos jours de nombreux types de supports de plus en plus miniaturisés permettent d'entreposer de l'information. La taille des disques durs se chiffre désormais en centaines de giga-octets et certains constructeurs d'ordinateurs personnels n'intègrent déjà plus de lecteurs de disquette sur leurs modèles d'ordinateurs. L'apparition des ordinateurs personnels (PC), la baisse des coûts entraînés par les évolutions techniques et l'apparition d'interfaces plus conviviales donnent la possibilité au grand public d'y avoir accès et de créer ses propres informations numériques. De nouveaux supports plus fiables apparaissent également depuis quelques années sur le marché. Leurs interfaces d'utilisation conviviales ou leur facilité de transport relativement

à la quantité d'information qu'ils peuvent contenir séduisent de plus en plus d'utilisateurs. Nous pouvons citer notamment l'apparition des clés *USB* (USB pour *Universal Serial Bus* est un bus qui permet de connecter des périphériques externes à un ordinateur, ce bus supporte les branchements et débranchements à chaud), des lecteurs portatifs de musique numérique comme le *iPod*, des caméras et appareils photos numériques de haute résolution et de capacité de stockage de plus en plus grande. Les évolutions matérielles concernent aussi des appareils utilisés tant sur le plan professionnel que sur le plan personnel comme les téléphones portables permettant d'échanger de différentes façons de l'information (messages textuels (*SMS*) ou multimedia (*MMS*), navigation sur Internet) et de stocker aussi de l'information.

Les formats de stockage d'information évoluent également. L'utilisation courante des formats comme le *MP3*[3], qui est un format de compression de fichiers audio facilite la portabilité et l'échange d'information et donc la réplication d'information.
La création des *autoroutes* de l'information est également une cause qui explique le phénomène de croissance. La mise en place d'applications telles que la messagerie électronique, le transfert de fichiers, la connexion à des machines distantes afin d'en utiliser les ressources a rapidement permis la connaissance par le grand public des possibilités d'un tel réseau. De nos jours, les universités, les organisations de tout type, les administrations, les entreprises ou les particuliers dans presque tous les pays du monde peuvent avoir de plus en plus facilement accès à ce réseau. Toutes ces structures génèrent donc de l'information, la traitent, et l'échangent.

Croissance du nombre de services électroniques

Les services électroniques sont directement liés aux évolutions technologiques. De nos jours, ces services touchent pratiquement tous les aspects de la vie professionnelle et personnelle. Le premier concerne la possibilité pour les utilisateurs de créer leur propres informations numériques au travers d'interfaces de plus en plus conviviales et les possibilités de mettre à disposition ces informations à travers les réseaux en utilisant les services web : messagerie électronique, forums de discussion, site personnel, *blogs*. Désormais, les différents outils peuvent également communiquer : envoi d'informations (un message *sms* par exemple) d'un ordinateur vers un téléphone portable et inversement, ce qui facilite davantage la création de nouvelles informations. L'information représente finalement aujourd'hui une mine que de multiples organisations transforment et structurent en connaissances à capitaliser pour faire évoluer l'entreprise. Cette capitalisation passe par la mise en œuvre d'outils et méthodes favorisant la structuration et la circulation des

[3]Le sigle MP3 est un raccourci de MPEG-1/2 Audio Layer 3, qui désigne la spécification sonore du standard MPEG1 (*Motion Picture Expert Group*)

informations stratégiques. L'intérêt des grands groupes est grandissant et se manifeste à travers des outils comme le portail d'entreprise ; ce type de porte d'entrée vers les informations de l'entreprise pour l'ensemble du personnel de l'entreprise et éventuellement ses partenaires existe également pour les organisations, certains corps de metiers mais aussi les universités. Pour ces dernières, l'utilisation des réseaux sert à mutualiser les ressources au travers de portail de ressources pédagogiques. Cette utilisation des réseaux a surtout donné lieu à une nouvelle forme d'apprentissage nommé *e-learning*, qui s'appuie en général sur une plate-forme informatique de gestion du contenu des formations et des parcours académiques.

La possibilité d'échanger de l'information de façon sécurisée sur les réseaux informatiques a transformé certains processus économiques en s'y intégrant et a donné lieu à de nouvelles formes d'entreprises électroniques ou *e-business*. Leur fonctionnement consiste à intégrer le réseau Internet dans le cadre général des activités de l'entreprise, pour la gestion de la relation avec la clientèle, pour les échanges avec des partenaires comme des fournisseurs ou des sous-traitants ou d'autres filiales de l'entreprise situé sur un emplacement géographique différent. Le fonctionnement de ce type d'entreprise concerne désormais tous les secteurs économiques, politiques, sociaux, etc.

Croissance du nombre d'utilisateurs

Les évolutions technologiques ainsi que l'offre grandissante de services électroniques entraînent une croissance fulgurante du nombre d'utilisateurs. Le besoin d'échanger rapidement et efficacement de l'information numérique devient plus courant et il accélère l'augmentation du nombre d'utilisateurs. D'après des chiffres récents de l'INSEE[4], un français sur quatre utilisait Internet au quotidien en 2005. 59% des personnes en France entre 16 et 74 ans ont déjà utilisé internet et 28% l'utilisent tous les jours ou presque. La quasi-totalité des jeunes, des cadres et des diplômés de l'enseignement supérieur sont aujourd'hui initiés à l'usage du web. Depuis 2001, en moyenne chaque année, trois millions de nouveaux utilisateurs goûtent à l'internet. La France occupe avec ces chiffres une position juste intermédiaire en terme d'usages d'internet, parmi les pays européens [52]. Aujourd'hui, plus de 43% des individus sont connectés à domicile [52] alors qu'en 1998, seul un ménage sur cinq disposait d'un ordinateur [44]. En 2004, les biens et services des technologies de l'information et de la communication ont également été des moteurs de la croissance. Les lieux de connexion vont des endroits publics (bibliothèque, cybercafé, lieu de travail ou d'étude) à des endroits privés (maison, chez des membres de la famille ou chez des amis).

[4]Institut National de la Statistique et des Études Économiques en France, l'étude utilisée dans cette section s'est appuyée sur des données allant jusqu'au mois d'Octobre 2005

TAB. 2.1 – Distribution des utilisateurs d'Internet par continent (en millions) ([98])

Région	**Utilisateurs (en millions)**
Afrique	6,31
Asie Pacifique	187,24
Europe	190,91
Middle East	5,12
Canada et USA	182,67
Amérique Latine	33,35

Les résultats de l'étude de *Lyman et al.* [98] donnent la répartition des utilisateurs d'Internet par continent dans le monde (tableau 2.1).

2.2.2 Etudes quantitatives

Il n'est pas aisé de quantifier le volume d'information numérique produit, ni d'avoir le nombre exact d'utilisateurs de ces informations. Des études ont toutefois réalisé des estimations dans différents domaines.

Dans les bibliothèques numériques

Depuis les années 90, des milliers de bibliothèques numériques de forme et de contenu très variables ont vu le jour. Un répertoire de bibliothèques sur le web maintenu à l'Université de Californie, Berkeley (*Libweb*[5]) liste couramment environ 7 100 pages de bibliothèques dans près de 115 pays. Des centaines de projets (internationaux, européens, nationaux, locaux) ont vu le jour dans de nombreux pays en ce qui concerne les bibliothèques numériques [135]. Nous pouvons citer entre autres *The Telematics for Libraries program of the European Commission*, *Candle (Controlled Access to Network Digital Libraries in Europe)*, *Decomate II (Delivery of Copyright Materials in Electronic Form)* comme projets financés par la Commission européenne. *Raitt* [119] liste de nombreuses initiatives nationales sur les bibliothèques numériques et des initiatives à l'échelle locale.

Sur le Web de surface et le Web invisible

La croissance d'information concerne fortement les deux zones du web que sont le *web visible* (pages statiques, pages disponibles pour le grand public) et le *web invisible* (un ensemble de

[5]http ://www.kb.nl/infolev/libweb/ ou http ://lists.webjunction.org/libweb/

bases d'information spécialisées accessibles via le web, des sites dynamiques) qui était estimé à 400 à 550 fois plus vaste que le *web visible* en 2001 [19]. Les études quantitatives concernant le Web sont délicates à résumer car les périodes concernées par les études se chevauchent souvent et certaines études s'appuient sur un type particulier d'information (articles de journaux, sites Web gouvernementaux et accessibles au grand public, informations créées par des particuliers, etc). En 2002, sur le web, la masse d'information correspondant aux pages statiques et aux pages calculées sans paramètres (par opposition aux pages produites dynamiquement à l'issue d'une demande *via* un formulaire par exemple) est estimée à 170 tera-octets. Une augmentation régulière de 30% par an pour ces pages a été constatée entre 1999 et 2002 [28]. Une étude estime la quantité d'information totale produite en 2002 à 5 exa-octets[6] dont 92% serait stockée de façon numérique. En 2001, le *web visible* est constitué d'approximativement $2,5$ millions de documents avec une vitesse de croissance de $7,3$ millions de pages par jour [19, 39]. La taille moyenne d'une page du *web visible* est dans l'intervalle [10 Ko , 20 Ko], ce qui donne entre 20 à 25 téra-octets pour le *web visible* (documents *HTML*), soit entre 10 et 20 téra-octets de contenu textuel ([93] utilise un facteur de $0,4$ pour passer de la taille en *HTML* à la taille du contenu textuel).

Messagerie électronique

Les messages électroniques font partie des moyens de communication les plus utilisés de nos jours dans un cadre personnel et/ou professionnel. Les estimations sur la quantité d'information véhiculée varient de 610 milliards à 1.100 milliards de messages envoyés au cours de l'année 2001. La taille moyenne d'un message mail étant de 18.500 octets, le flux d'information est de l'ordre de 11.285 à 20.350 téra-octets, mais il faut préciser que tout ce flux n'est pas stocké de façon permanente [98].

Les listes de diffusion peuvent être vues comme une sous-catégorie des messages électroniques ; *Listserv*[7], un des gestionnaires de listes de diffusion les plus fréquemment utilisés sert à envoyer environ 30 millions de messages par jour dans environ 150.000 listes. Un échantillon de listes de diffusion a montré que 30% d'entre elles passent par *Listserv* ; en se basant sur ces statistiques, on peut estimer le nombre total de listes de diffusion à environ $36,5$ milliards par an soit un volume cumulé de 675 Tera-octets [98].

[6] 1 exa $\approx 10^{18}$

[7] http ://www.lsoft.com/catalist.html

TAB. 2.2 – Taille de l'Internet en Tera-octets. ([98]

Catégorie	nombre de Tera-octets (année 2002)
web visible	167
web invisible	91 850
E-mails	440 606
Instant messaging	274
TOTAL	532 897

Autres services web : ftp, IRC

Peu de données existent en ce qui concerne le *FTP*(*File Transfert Protocol*) et de plus en plus d'archives accessibles sous *FTP* le deviennent aussi sur *HTTP* (*HyperText Transfert Protocol*) ; [98] a fourni l'exemple de deux serveurs ftp qui contiennent environ 0, 412 téra-octets de données (*ftp.cdrom.com* et *ftp.freesoftware.com*).

Le *Internet Relay Chat* (IRC et en français *discussion relayé par internet*) et les *messaging services* (services de messagerie comme MMS pour *Multimedia Messaging service*) représentent plus un flux qu'un stock d'information. *Liszt.com* qui est un des plus gros répertoire de canaux IRC possède 37 750 canaux sur 27 réseaux, avec 150 000 utilisateurs, chacun ayant une très bonne vitesse de saisie [98].
Les travaux de *Lyman et al.* [98] ont fourni en 2003 une étude sur la quantité d'information produite. Pour ce qui est du web, le tableau 2.2 donne le résumé de leur résultats.

Les *blogs* et bibliothèques numériques personnelles

Les données personnelles forment une part non négligeable dans la masse d'information actuelle. Les espaces de production et publication d'information pour des particuliers sont de plus en plus nombreux. Le phénomène des *blogs* (contraction de *web log* ou carnet de bord Web) est récent mais son envergure est sans précédent ; les travaux qui étudient leur croissance se mettent à jour très régulièrement à cause des changements rapides qui s'opèrent au niveau des chiffres concernant leur croissance. Un *blog* est un site web sur lequel une ou plusieurs personnes s'expriment de façon libre, sur la base d'une certaine périodicité. Les résultats publiés en 2005 par *David Sifrys* qui s'appuient sur des études de *Technocrati*[2] montrent que la taille de la *blogosphère*[8] double désormais tous les six mois environ. 75 000 nouveaux *blogs* sont crées chaque jour, soit environ un nouveau *blog* chaque seconde. Désormais, il existe des systèmes de RI dédiés

[8]Ensemble des *blogs* disponibles sur le Web.

TAB. 2.3 – Production mondiale d'information originale en tera-octets, si stockée de façon numérique. L'estimation supérieure suppose que l'information rendue est numérique et l'estimation inférieure suppose une compression de l'information. BS :Borne supérieure, BI :Borne inférieure. Le pourcentage de changement est donnée par rapport à la production d'information en 2002 [98]

Support	$1999-2000$ **BS**	$1999-2000$ **BI**	**% de changement BS**
Papier	1.200	240	36%
Film	431.690	58.209	-3%
Magnétique	2.779.760	2.073.760	87%
Optique	81	29	28%
TOTAL	3.212.731	2.132.238	$74,5\%$

à la *blogosphère*.

Les travaux de *Lyman* [98] donnent un aperçu global de la production d'informations, en classant ces informations suivant leurs supports possibles (papier, film, magnétique, optique). Le tableau 2.3 est tiré de leurs travaux.

2.2.3 Les moteurs de recherche face au passage à l'échelle

Une étude récente évalue les résultats de RI sur un certain nombre de moteurs de recherches leaders sur le web en utilisant un ensemble de requêtes-utilisateur *réelles* sélectionnées de façon aléatoire. Cette étude montre que seulement $3,2\%$ des premières pages retournées sont identiques pour les trois meilleurs moteurs de recherche pour une requête donnée et les moteurs de recherche ont en commun un de leurs cinq premiers documents dans moins de 20% de cas [41]. En effet, la taille du web couplée à son dynamisme et à l'ambiguïté (la subjectivité) des requêtes-utilisateurs font qu'il est difficile pour un moteur de recherche de fournir l'information la plus actualisée en temps réel. Une étude de *Ding et Marchionini*[158] montrait en 1996 qu'il y a une partie non négligeable du web qui n'est indexée ou couverte par aucun des moteurs de recherche existants. Cette étude pointait déjà le peu de chevauchement entre les résultats de moteurs de recherche différents interrogées avec les mêmes requêtes. *Lawrence and Giles* [92] en 1998 ont montré également qu'un moteur de recherche n'indexait pas plus de 16% de tous les sites web (cité dans [41]). Ces résultats sont en conformité avec des études moins récentes dans le monde académique. Les travaux de *McGill et al.* en 1979 [104] avaient également détecté peu de chevauchement entre des ensembles de documents retournés pour un même besoin d'information avec différentes techniques ou différentes formulations ; les travaux de *Saracevic et Kantor* en 1968 [136] portant

également sur différentes formulations d'un même besoin d'information ont eu des résultats similaires. Les travaux de *Katzer et al* [85] notaient le même phénomène en utilisant différentes représentations de documents (titre, résumé, etc) en 1982. Ainsi, l'étude de [41] de 2005 a révélé que le pourcentage de résultats uniques à un moteur de recherche est de plus de 84% et le pourcentage de résultats communs aux 4 moteurs de recherche que sont *Google*[9], *Yahoo !*[10], *MSN*[11] *et Ask Jeeves*[12] est de $1,1\%$. Ainsi en menant une recherche uniquement sur *Google*, un utilisateur peut *rater* jusqu'à 70% des meilleures premières pages de résultats possibles (*web's best first page search results*). Une des solutions apportées à ce problème de peu de chevauchement entre différents moteurs de recherche est l'utilisation de méta-moteurs de recherche[13]. Des travaux de recherche en RI ont utilisé différentes techniques de fusion pour combiner les résultats issus de différents modèles de RI ou différentes formulations d'un même besoin d'information. On peut citer les travaux de *Turtle et Croft* [62] sur le système *INQUERY*, ceux de *Fox et Shaw* [49], ceux de *Belkin et al.* [16] et ceux de *Lee* [94] qui utilisent différents schémas de pondération pour retourner plusieurs ensembles de documents, face à une requête ; la combinaison de ces ensembles fournit des améliorations de performance.

Face au passage à l'échelle, les moteurs de recherche font donc ce qu'ils peuvent. Leurs différentes techniques prennent souvent en compte la croissance des informations de façon plus ou moins explicite. Nous présentons quelques techniques utilisées dans divers types de systèmes d'information, qui par leur nature participent à la prise en compte du passage à l'échelle. Nous mettons ensuite en exergue les liens *naturels* entre certains axes de la RI et le passage à l'échelle.

2.2.4 Mise en œuvre de systèmes d'informations et passage à l'échelle

Des techniques en systèmes d'information en lien avec le passage à l'échelle

La prise en compte du passage à l'échelle se fait de façon implicite par certaines techniques de mise en œuvre de systèmes d'information. Ces techniques sont basées généralement sur des modèles qui consistent à *diviser pour régner* ou à *mieux connaître son adversaire pour mieux le combattre.*

Architectures distribuées et parallèles L'utilisation du calcul parallèle est nécessaire pour diminuer les temps de traitement et augmenter l'espace mémoire utilisable durant les traitements que font les ordinateurs. Le but principal de cette technique est de prendre en

[9] http ://www.google.com
[10] http ://www.search.yahoo.com
[11] http ://search.msn.fr/
[12] http ://www.ask.com
[13] exemple : http ://www.dogpile.com, http ://www.seek.fr

compte de façon plus aisée les traitements nombreux ou des traitements longs et pouvant se faire de façon simultanée. Quand la quantité d'information à traiter augmente, l'utilisation du calcul parallèle est une arme que de nombreux travaux utilisent dans plusieurs domaines et notamment en RI.

L'objectif des architectures distribuées est de décentraliser la gestion des informations et de répartir les traitements sur différents ordinateurs. Ce type d'architecture est couramment utilisé par les moteurs de RI sur le web et permet la prise en compte de l'augmentation en quantité des informations.

Utilisation des méta-données Une méta-donnée est une donnée sur une donnée. Plus précisément, on peut définir les méta-données d'une ressource comme un ensemble d'informations la décrivant et utiles pour son utilisation. L'emploi de ce concept qui provient à l'origine du monde des bibliothèques et de la documentation s'est répandu avec l'avènement de l'édition numérique et de l'informatique. Ainsi, les différents langages d'édition sur le Web par exemple prévoient l'insertion de méta-données internes dans l'entête des documents. Bien que peu utilisées par les moteurs de recherche, des efforts de normalisation existent pour ces meta-données mais aussi pour les documents dans le but d'aboutir au Web sémantique, un web plus exploitable de façon automatique et intelligente. Des travaux de RI ont opté pour la prise en compte des méta-données dans leur démarche. La valeur ajoutée des meta-données serait de permettre d'écarter les informations parasites (le bruit documentaire) et de réduire simultanément les silences documentaires (les informations pertinentes existantes mais non rapportées), ces deux problèmes étant particulièrement évidents quand on traite des grandes quantités d'informations.

Utilisation des *ontologies* en RI En informatique, le mot *ontologie* est parfois utilisé pour désigner un ensemble structuré de concepts. Les concepts sont organisés dans un graphe dont les relations peuvent être des relations sémantiques, des relations de composition et d'héritage (au sens objet). Parfois, ce type de graphes est aussi nommé thésaurus, et l'appellation ontologie est remise en cause par certains chercheurs. En RI, l'organisation des informations en graphes est une méthode adoptée par plusieurs travaux. Quand la masse d'information est volumineuse, la structuration des informations devrait aider à guider le système de recherche vers les informations en lien avec le besoin de l'utilisateur et élaguer les informations moins intéressantes.

Axes de RI liés « naturellement »au passage à l'échelle

Classification de documents La classification des documents consiste à attribuer à tout document une ou plusieurs catégories, en se basant sur son contenu. Cette attribution peut être supervisée (un mécanisme extérieur comme l'intervention humaine fournit une classification correcte qui va être utilisée comme base de départ) ou non supervisée (la classification est entièrement réalisée automatiquement sans intervention externe au système qui classifie). Dans l'une ou l'autre des méthodes, le fait d'attribuer une catégorie à un document a pour objectif d'y avoir accès plus aisément ultérieurement. Il s'agit de regrouper les textes similaires, c'est-à-dire thématiquement proches, au sein d'un même ensemble. L'intérêt d'une telle démarche est d'organiser les connaissances de façon à pouvoir effectuer, par la suite, une recherche ou une extraction d'information efficace. C'est une approche qui compte plusieurs travaux de recherche et qui facilite par son essence (*diviser pour mieux accéder*) la prise en compte du volume croissant d'information.

RI multilingue et crosslingue De nombreuses langues sont représentées dans l'ensemble des informations à disposition de nos jours. La prise en compte de ces diverses langues lors du processus de RI peut se faire soit par traduction du besoin d'information de l'utilisateur soit par traduction des documents au sein desquels la recherche s'effectue. La première approche est celle qui a été le plus souvent utilisée mais des travaux s'intéressent désormais à fournir des représentations de documents pouvant permettre leur traduction/transposition aisée en d'autres langues. Il s'agit de faire émerger dans plusieurs langues des structures de sémantiques semblables afin que la recherche multilingue se fasse au niveau des termes d'indexation et non plus par traduction des requêtes[14]. La prise en compte de plusieurs langues lors de la recherche d'information facilite le travail de l'utilisateur face au volume d'information très grand et lui permet de mener sa recherche dans la langue de son souhait, tout en étant assuré d'obtenir tous les résultats existants, quelles que soient leurs langues d'écriture. Les travaux de la campagne d'évaluation *CLEF* [118] participent à cette dynamique.

RI interactive et retour de pertinence L'objectif premier des systèmes de RI est de satisfaire le besoin d'information de l'utilisateur. Ce but situe au cœur du problème la nécessité d'intégrer au modèle de RI les usagers du système notamment en mettant en œuvre une pertinence-système la plus proche possible de la pertinence-utilisateur. Dans ce contexte où l'interaction homme-système est primordiale, des travaux étendent la notion de modèle de

[14] Axes de travail du laboratoire IMAG : `http://www-mrim.imag.fr/presentation/multilingue.php`

RI (notion sur laquelle nous reviendrons en section 2.5.1), de même que les réalisations qui en sont dérivées, à la prise en compte plus large de la RI interactive. Il devient encore plus urgent dans un contexte de grands volumes d'informations de fournir des moyens aux usagers pour spécifier le plus clairement possible leur besoin d'information, en leur permettant d'interagir avec les systèmes de recherche.

Résumé de textes Pour permettre d'accéder à l'information essentielle des textes, des outils utilisant la linguistique sont mis en œuvre pour produire automatiquement des résumés de texte. De nombreux travaux s'intéressent à cet axe de recherche et permettent de présenter les phrases les plus informatives en les représentant de diverses façons (par exemple par des couleurs plus ou moins dégradées en fonction de l'importance de la phrase). Lors du processus de RI, le volume énorme d'informations actuel peut être réduit considérablement par ces techniques et permettre d'élaguer les textes peu intéressants pour les utilisateurs, et pousser plus loin la recherche uniquement dans les textes dont les résumés intéressent les utilisateurs. Il est à noter toutefois que les techniques actuelles de résumés automatiques de texte sont encore loin de produire des résumés similaires à des résumés réalisés par des humains.

Méta-RI et Fusion de données Les méta-moteurs sont des outils qui, pour une même requête, interrogent plusieurs moteurs de façon simultanée, rapatrient leurs résultats, les synthétisent et proposent un récapitulatif des réponses données. Le principe de fonctionnement est séduisant quand on sait que sur le Web par exemple, aucun moteur de recherche n'indexe la totalité des informations disponibles. Cette dernière remarque induit un taux de chevauchement très faible entre les résultats obtenus pour un même besoin d'information exprimés sur différents moteurs de recherche comme décrit dans la section 2.2.3. Ce faible taux diminue en même temps que la masse d'information grandit. Les méta-moteurs devraient donc permettre de pallier (au moins partiellement) à ce problème et à retrouver les informations les plus intéressantes pour un besoin d'information. Toutefois, la mise en place des fonctionnalités avancées dans le cadre d'une recherche simultanée sur plusieurs moteurs est loin d'être aisée, voire tout simplement impossible. De plus, les méta-moteurs font la synthèse de résultats fournis par plusieurs moteurs différents, classant chacun leurs résultats de façons différentes, sans forcément utiliser les mêmes critères de pertinence. De nombreux travaux s'intéressent à améliorer les techniques de fusion de résultats provenant de systèmes de recherche différents et tendent à obtenir des résultats encourageants (*Manmatha et al.* [100], *Lee* [95], *Croft* [36], *Aslam et Montague* [9], *Manmatha* [101]).

Tous ces aspects nous montrent que l'accès à l'information est prise en compte de différentes façons. Toutefois, les étapes de base qui constituent le processus de RI sont les suivantes : tout d'abord, la phase de collecte des informations formant l'espace de recherche, ensuite la structuration de cet espace de recherche par indexation et enfin la prise en compte du besoin d'information de l'utilisateur. Le système de recherche réalise ensuite l'appariement entre le besoin d'information de l'utilisateur et les éléments contenus dans l'index des informations collectées : les éléments qui sont les plus intéressants d'après le système de RI sont alors retournés à l'utilisateur et ceci doit se faire sous une forme accessible, compréhensible et exploitable pour l'utilisateur. Nous nous intéressons dans les sections suivantes à la prise en compte du passage à l'échelle sur chacune de ces phases.

2.3 Mise en place de l'espace de recherche : prise en compte du passage à l'échelle

Le processus de RI consiste à fournir, en réponse à une demande de l'utilisateur (couramment appelée *requête*) les *documents* qui répondent au mieux à son besoin d'information (documents pertinents). Cet objectif introduit des concepts sur lesquels nous reviendrons au long des différents chapitres composant cette thèse.

La notion de *document* peut représenter dans la réalité différentes unités d'information : une page HTML (sur le Web par exemple), un paragraphe de texte, un *doxel*[15] (recherche dans les documents structurés de type XML par exemple), quelques phrases d'un texte (résumé de textes), ... mais dans tous les cas, c'est le résultat fourni en réponse à un besoin d'information par un système de RI.

La notion de *requête* est quant à elle utilisée en RI pour désigner la demande d'information de l'utilisateur ; cette demande peut suivre un formalisme plus ou moins rigoureux selon les systèmes de RI.

La notion de *pertinence* est également une notion sur laquelle nous reviendrons amplement dans le chapitre 3. C'est un concept cognitif complexe et subjectif puisque directement lié à des notions comme la satisfaction de l'utilisateur, l'utilité d'un document retourné, qui sont par nature des notions subjectives.

Ces concepts importants étant introduits, nous nous intéressons à la première phase d'un processus de RI qui est la constitution de l'ensemble des informations au sein duquel va s'effectuer

[15]Le néologisme *doxel* est un raccourci de *document element* en Anglais et désigne une unité d'information par analogie avec le pixel, l'unité graphique élémentaire.

la recherche. Ceci peut se faire de différentes façons suivant les contraintes de l'environnement de recherche.

2.3.1 Construction de collections volumineuses

Constituer l'espace de recherche sur le Web

Dans des environnements dynamiques comme le Web, deux options sont possibles pour la mise en place de l'espace de recherche : soit les créateurs de pages Web les soumettent au moteur de recherche, soit le moteur de recherche parcourt à une fréquence donnée une partie du Web et récupère un certain nombre de pages web qui constitueront l'espace au sein duquel va se faire la recherche. Dans ce dernier cas, la collecte de documents se fait par un *spider* ou logiciel robot d'indexation : tout d'abord le *spider* visite une page Web, la charge, l'archive et suit les liens partant de cette page pour visiter d'autres pages. Le *spider* retourne sur le site en question avec une fréquence (mensuelle, bimensuelle par exemple) pour recueillir les changements éventuels. Tout ce que le *spider* trouve est mis dans l'index ; si une page a changé, alors l'index est mis à jour. Une page qui est *spidered* (archivée) mais non encore indexée ne peut pas encore faire partie des résultats de recherche [41], [39]. Le dynamisme des collections du Web rend toutefois peu aisée la possibilité de reproduire des résultats de RI. Pour la communauté de recherche en RI, il s'est donc avéré nécessaire de mettre en place des ensembles d'informations statiques au sein desquels il est possible de tester différents systèmes de recherche sur une même base d'informations et de façon reproductible.

Les collections de test de Cranfield à nos jours

La RI a une longue tradition expérimentale qui utilise des collections pour faire avancer le domaine et dont témoigne des travaux datant des années 1960 (*Cleverdon et al.* [33], de *Salton* [126] ou de *Sparck Jones*[81]). Les collections dites de test utilisées pour la recherche se composent d'un ensemble de documents (corpus), d'un ensemble de besoins d'informations et d'un ensemble de jugements de pertinence sur les documents par rapport à ces besoins d'information. La collection de test idéale serait celle au sein de laquelle tous les documents ont un jugement de pertinence en regard de tous les besoins d'informations. Les premières collections ont été construites dans les années 1960 suivant ce principe. Leur taille réduite (par exemple celle de *Cranfield* créée en 1960 est composée de 1 400 documents avec 225 besoins d'information) permettait de juger la pertinence de tous les documents constituant l'espace de recherche. Les premières collections qui ont été effectivement utilisés étaient constitués d'au plus quelques milliers de documents :

la collection *CACM* (pour *Communications of the ACM*[16], collection de titres et de résumés provenant du journal de l'*ACM* entre 1958 et 1979, 3204 documents) décrite par *Baeza-Yates et Ribeiro-Neto* sur la base de l'étude de *Fox*[48], la collection *ISI*[17] créée par *Small* [142] (1460 documents). D'autres collections de test liées à des domaines particuliers ont également vu le jour (*NLM* pour le domaine médical (3078 documents), *NPL* pour l'ingénierie électrique (11 429 documents) ; chacune de ces collections est détaillée dans l'ouvrage de *Baeza-Yates et Ribeiro-Neto* [10].

La première collection créée par le *National Institute of Standards and Technology (NIST)* pour la Conférence TREC, *Text REtrieval Conference* voit le jour en 1992 et compte 2 Giga-octets d'articles de journaux du *Wall Street Journal* et de documents du gouvernement. A cette première campagne, 25 groupes ont participé à l'une ou l'autre des 2 pistes de recherche. En 2004, cette même conférence regroupait 103 groupes participants de 20 pays et 7 pistes de recherche, ce qui démontre l'agrandissement de la communauté de recherche s'intéressant aux problèmes soulevés par la RI.

L'initiative *CLIR* de *TREC* qui s'intéresse particulièrement à la RI multilingue et crosslingue (TREC-6 à TREC-8) a depuis été reprise par la conférence *Cross Language Evaluation Forum* [118].

L'objectif de ces collections de test est de fournir un environnement pour tester, évaluer et comparer différents systèmes de RI de façon reproductible. Le tableau 2.4 donne l'évolution du nombre de participants dans les différentes pistes des campagnes *TREC* depuis sa création. D'autres campagnes d'évaluation de systèmes de RI ont vu le jour depuis. Parmi ces initiatives, nous pouvons citer le projet *NTCIR NACCSIS-NII Test Collection for Information Retrieval*[18] dont la première campagne d'évaluation a eu lien entre Novembre 1998 et Septembre 1999 et qui a pour but l'évaluation des systèmes de RI pour le texte en japonais notamment [8].

Pour l'évaluation et la comparaison des systèmes de RI dans les collections de documents structurés (XML principalement), la campagne *INEX (INitiative for the Evaluation of XML retrieval)*[3] a été mise en place depuis 2000.

Des collections dédiées à des domaines spécifiques ont également vu le jour ; c'est le cas par exemple de la collection *OHSUMED* pour le domaine médical [75].

Nous donnons ici des exemples de documents extraits des collections de test de la campagne *TREC9*. La figure 2.2 présente un document contenant du texte et au sein duquel les balises

[16] Association of Computer Machinery

[17] Institute of Scientific Information

[18] National Center for Science Information Systems (NACSIS) connu depuis Avril 2000 comme le National Institute of Informatics (NII)

TAB. 2.4 – Quelques pistes de *TREC* avec le nombre de participants par année et quelques grandes collections.

Pistes	TREC													
	92	93	94	95	96	97	98	99	00	01	02	03	04	05
Total	22	31	33	36	38	51	56	66	69	87	93	93	103	
Ad Hoc	18	24	26	23	28	31	42	41						
Routing	16	25	25	15	16	21								
Interactive			3	11	2	9	8	7	6	6	6			
Filtering				4	7	10	12	14	15	19	21			
NLP					4	2								
Speech						13	10	10	3					
CrossLanguage						13	9	13	16	10	9			
High Precision						5	4							
VLC						7(20Go)	6(VLC2)							
QA								20	28	36	34	33	28	
Web								17(10Go)	23(VLC2,wt10g)	30	23	27	18	
Novelty											13	14	14	
Genomics												29	33	41
HARD												14	16	16
Robust												16	14	17
Entreprise														5,7Go
Spam														12
Terabyte													17(426Go)	(426Go)

FIG. 2.2 – Exemple de document issu de la collection WT10G de TREC9. Les balises HTML ont été enlevés.

<DOC>
<DOCNO>WTX001-B06-3</DOCNO>
ABOUT THE RAC The RAC is seen as one of the most progressive and dynamic motorists' organisations. We're certainly interested in investing in new technology if it means we can provide our Members with an even better service. Take our state of the art response systems, for example - thanks to our sophisticated network, RAC patrols can reach most breakdowns in just 40 minutes ! It's very important to us that we can offer our Members support and assistance whatever problems they may encounter on the road. We have recently changed our service structure to give you the choice of three new improved levels of cover - Rescue ; Rescue and Recovery ; Reflex. Each level covers you as a passenger or a driver, no matter what vehicle you're travelling in. We also offer a range of optional extras which you can choose to add to your level of cover. That way you make sure you benefit from the level of protection exactly suited to your driving needs. We re-invest our profits every year, so that we can continue to offer our Members enhanced services and new products, and provide a carefully balanced portfolio of benefits. As a result, when you join the RAC you'll find that we can give you so much more than roadside assistance ! Bradley Stoke Super Centre HOT OFF THE NET Please feel free to comment on our Web Site The Benefits of Membership - Off the Road Services - How to Join
</DOC>

HTML ont été enlevées ; par contre la figure 2.3 présente un document avec les balises *HTML*, mais ce document ne contient finalement qu'une *URL* et la question de la validité d'un tel document peut se poser (à l'origine, il pouvait contenir par exemple des images).

La croissance d'information entraîne le besoin de passer à des collections de test de plus en plus volumineuses, pour fournir un environnement de test réaliste. Toutefois, de par leur taille, il n'est plus possible de fournir des jugements de pertinence pour tous les documents comme c'était le cas pour les petites collections (*Cranfield*). Aussi, la technique du *pooling* a été conçue pour leur construction et décrite dans les rapports de *Spärck Jones et Bates*[82] en 1977 et de *Gilbert et Jones* en 1979 [56]. Cette technique est détaillée dans le chapitre portant sur l'évaluation (chapitre 3). En effet, la méthodologie d'évaluation des systèmes de RI s'appuie sur les collections de test et est donc fortement liée à la façon de les mettre en place.

Toutefois, les ressources ne permettent pas toujours de travailler sur des collections de grande

FIG. 2.3 – Exemple de document issu de la collection WT10G de TREC9 avec les balises HTML

<DOC>
<DOCNO>WTX001-B01-9</DOCNO>
<DOCOLDNO>IA001-000000-B006-167</DOCOLDNO>
<DOCHDR> http ://sd48.mountain-inter.net :80/hss/teachers/Perdue.html 204.244.59.33 19970101012526 text/html 435 HTTP/1.0 200 OK Date : Wed, 01 Jan 1997 01 :15 :43 GMT Server : Apache/1.0.3 Content-type : text/html Content-length : 265 Last-modified : Mon, 25 Nov 1996 05 :31 :23 GMT </DOCHDR>
</ref="teachers.html">Back to Teachers' Home Page Y></HTML>
</DOC>

taille. De plus, ces collections elles-mêmes ne constituent qu'une partie de l'ensemble des informations accessibles par exemple aux moteurs de recherche sur le Web. Pour étudier le comportement des systèmes sur un type de collection donnée par exemple des collections contenant des hyperliens de façon reproductible, il est nécessaire de créer à un instant donné un représentant du type concerné de collection pour mener les études. De même, pour étudier la façon dont se comportera un système au fur et à mesure que la collection sur laquelle il travaille augmente en quantité d'information, il est nécessaire de pouvoir disposer de collections de tailles différentes. Il est également nécessaire dans ce cas de disposer d'une représentation de la croissance de collections.

2.3.2 Prise en compte du passage à l'échelle

Des travaux se sont intéressés aux différentes façons de faciliter la mise en place de collections volumineuses et à leur utilisation pour étudier le passage à l'échelle. Dans le cas de SRI utilisant plusieurs sources distribuées *Chowdury et al.* [29] (ou *Frieder et al.* [50]) facilitent le passage à l'échelle dans la taille de corpus en détectant et en réduisant la duplication des documents identiques provenant de sources différentes. En effet, il ne serait d'aucun apport de fournir le même document plusieurs fois en réponse à une demande de l'utilisateur. De plus, la duplication peut affecter les performances de SRI se basant sur la fréquence documentaire[19] des termes par exemple.

Une des façons de prendre en compte l'accroissement en taille des collections est de segmenter ces collections pour éviter l'accroissement des temps de recherche par exemple. Le défi est d'identifier la base sur laquelle cette segmentation va être faite. Elle peut se faire à base de questions

[19] La fréquence documentaire est une fonction du nombre de documents au sein desquels le terme apparaît. Elle est utilisée comme base du calcul du facteur couramment nommé idf pour *inverse document frequency*.

auxquelles l'utilisateur répond ou d'un profil établi grâce à un historique des recherches [112], à base de méta-données portant sur le besoin en information de l'utilisateur et/ou sur la nature ou l'usage des documents [28]. Sur le Web, une solution est l'utilisation des noms de domaines *(.com, .edu, etc.)*.

Lorsque cette segmentation concerne les collections de test, et que l'on souhaite étudier l'impact de la taille sur les performances de SRI, il ne suffit pas que les tailles des collections soient croissantes pour que l'on puisse tirer des conclusions sur les performances du système. Si les collections sont constituées de façon trop différente, il sera difficile de dissocier dans les résultats obtenus la part due à l'influence de la taille de la collection et la part due aux (contenu des) collections elles-mêmes. Il est nécessaire de connaître et de contrôler le biais induit par ce qui constitue les collections. Une des façons de le faire est de créer des échantillons *similaires* de collections et de taille croissante.

L'échantillonnage

Un échantillon est une fraction représentative d'une population. En économétrie et théorie des sondages, on désigne par échantillon un ensemble d'individus extraits d'une population initiale. Cet ensemble doit obéir à un certain nombre de règles afin de s'assurer de l'homogénéité et de la représentativité de ce groupe. Le point central de l'échantillonnage est la *représentativité* des échantillons construits. La statistique s'appuie également sur la notion d'échantillon puisqu'elle a pour but de faire émerger des propriétés d'un ensemble de variables connues uniquement à travers quelques unes de ses réalisations (qui constituent un échantillon de données). En théorie du signal, on parle d'échantillonnage pour qualifier la transformation d'un signal analogique (continu) en signal numérique (discret), en capturant des valeurs à intervalle de temps régulier (ici temps est à prendre au sens large et s'applique à tout signal). Des travaux antérieurs en recherche d'information bâtissent des sous-collections pour les études et parlent d'échantillons de collections, alors que d'autres travaux parlent simplement de sous-collections.

Etudier la distribution des scores de pertinence pour prédire le comportement des modèles face au passage à l'échelle

Les travaux de *Manmatha et al.* [100] ont étudié la forme que prend la distribution des scores de pertinence (*Retrieval Status Values* ou *RSV*) des documents pertinents (resp. non pertinents) pour différents modèles de RI. *Hawking et Robertson* [67] appliquent la théorie de détection de signal à la RI en utilisant différentes formes de distribution de score de documents pertinents et de score de documents non pertinents. Cette théorie leur permet d'obtenir pour chaque forme de

distribution de score les performances attendues du système en fonction du nombre de documents considérés.

Créer différentes sous-collections de tailles croissantes

Dans la partie expérimentale de leurs travaux utilisant la détection du signal et la distribution des scores de pertinence des documents, *Hawking et Robertson* [67] ont constitué trois types de sous-collections [20] de la collection entière :

- les sous-collections dites uniformes : on crée n échantillons primaires de taille égale. Ainsi les sous-collections composées de taille $2/n, 3/n, \cdots, (n-1)/n$ sont constituées en composant les n échantillons comme le montre l'exemple suivant pour un échantillon de taille $3/7$: on constitue tout d'abord 7 échantillons primaires disjoints numérotés $0, 1, 2, 3, 4, 5, 6$ et dont l'union est la collection initiale. On crée ensuite 7 échantillons composites comme suit : $(0, 1, 2), (1, 2, 3), (2, 3, 4), (3, 4, 5), (4, 5, 6), (5, 6, 0)$ et $(6, 0, 1)$. Les tests se font sur chacune de ces sous-collections composées et le résultat reporté pour l'échantillon de taille $3/7$ correspond à une moyenne des résultats sur toutes ces sous-collections. Ainsi, les mesures moyennes reportées pour chaque taille de sous-collections prennent en compte toutes les données de la collection initiale.
- les sous-collections répliquées : On prend les échantillons primaires de taille $1/10$ (par exemple) de la collection entière et on les réplique un nombre de fois voulu pour créer des sous-collections d'une certaine taille. Pour chaque taille de sous-collection, la valeur de performance retenue au final est une combinaison des valeurs de performance obtenues pour les différents échantillons de cette taille.
  Exemple : $(0), (0, 0), ..., (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, $(9), (9, 9), ..., (9, 9, 9, 9, 9, 9, 9, 9, 9, 9)$
- les sous-collections dites *biaisées* : Les données de *TREC-6* sont subdivisées en 5 sous-collections disjointes suivant leur origine. On fait de l'ensemble de documents de chacune des sources un échantillon, mais ces ensembles diffèrent en nombre de documents et en taille (de $235,4$ Mo à $564,1$ Mo) et leurs documents n'ont pas les mêmes probabilités de pertinence.

Les méthodes de *Hawking et Robertson* [67] de construction de collection de taille ont pour objectif de fournir un cadre d'étude de l'influence du passage à l'échelle sur les performances des systèmes de RI.

Les résultats obtenus en utilisant ces différents types d'échantillons de collections sont présentés

[20] Les méthodes de construction de ces sous-collections rendent discutable l'appellation d'échantillon, en tout cas sur le plan statistique

au chapitre 3.

Les travaux de *Beigbeder et Mercier* [15] se sont intéressés à l'évolution des distributions des valeurs de la fréquence des termes et de la fréquence documentaire traditionnellement utilisées dans le modèle vectoriel[21] de recherche d'information en fonction du nombre de documents indexés. Ces auteurs construisent des sous-collections (partant d'une collection initiale)[22] dont la taille augmente d'un facteur 10 à chaque fois. Il n'y a pas de critère pour le choix des documents à ajouter d'une sous-collection à la sous-collection suivante, c'est simplement l'ordre *natif* de la collection complète qui est utilisé.

Créer une sous-collection similaire à un ensemble volumineux donné

Gurrin et Smeaton [61] se sont intéressés à la réplication de la structure des hyper-liens du Web sur une collection de taille réduite. Ils analysent la structure des liens sur le Web et proposent des techniques pour la mise en place de collections au sein desquelles on retrouverait la même structure de liens. Ces collections statiques (à l'inverse du Web) permettraient des études reproductibles et l'évaluation de SRI utilisant les hyper-liens. Ceci concerne l'*échantillonnage* des collections. Il s'agit d'abord de déterminer les propriétés d'une collection volumineuse (des caractéristiques intéressantes pour le type d'étude que l'on souhaite réaliser par la suite), ensuite de construire une collection de taille réduite ayant les mêmes propriétés et de réaliser des études sur la collection réduite (ce qui devrait être plus aisée que sur la collection entière) et enfin (et surtout) de pouvoir reporter sur la collection entière les résultats obtenus sur cette collection réduite.

Simuler un phénomène lié à la croissance d'information

La simulation consiste à essayer de prédire ou de reproduire un comportement. La mise en place de collections pour simuler un phénomène donné quand la taille de collection augmente est une technique qui a été utilisée par des travaux antérieurs. L'incomplétude des jugements de pertinence augmente avec la taille de collection. Pour la simuler et pouvoir étudier son impact, *Buckley et Voorhees* [24] construisent, à partir d'une collection de départ supposée avoir des jugements de pertinence *complets*, une succession de collections sur lesquelles les jugements de pertinence sont de moins en moins nombreux. Les expérimentations sur ces ensembles de col-

[21] Plus de détails sur ce modèle en section 2.5.1.

[22] Les documents de la collection initiale sont extraits des pages HTML récoltées sur des sites de domaines géographiques francophones en décembre 2000. http ://www-mrim.imag.fr/membres/mathias.gery/Robot/WebFr4_01_12_2000/ domaines.html

lections montrent que les métriques classiques de RI ne sont pas robustes face à l'incomplétude. Aussi, une nouvelle métrique est proposée qui d'après ces travaux fait mieux face au problème d'incomplétude.

Nous proposons dans nos travaux une méthodologie de construction de sous-collections (de tailles croissantes) *similaires* à une collection initiale pour simuler la croissance de collection . Cette méthodologie est dans un second temps renforcé par les techniques connues et utilisées en statistique pour créer des échantillons d'un ensemble de données de façon à ce que les résultats obtenus sur ces échantillons puissent être valides statistiquement.

2.4 Passage à l'échelle pour l'Indexation

2.4.1 Création d'index

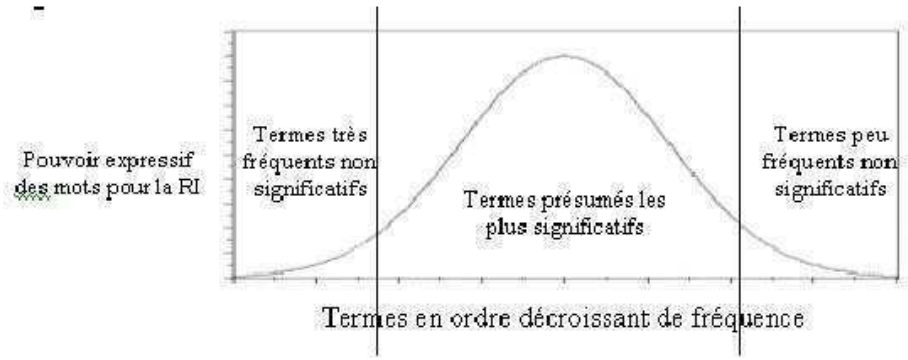
La mise en place de collections volumineuses pose des problèmes liés fondamentalement à l'espace de représentation. Il est nécessaire d'organiser l'ensemble des informations contenues dans la collection de façon à fournir efficacement et rapidement une réponse en cas de demande de l'utilisateur. Ainsi, les traitements préalables à la recherche peuvent être réalisés avant toute demande de l'utilisateur.

Cette étape, nommée *indexation* est préalable à la prise en compte des demandes des utilisateurs. Créer un index sur un ensemble d'informations c'est fournir une façon de localiser une information donnée dans l'ensemble. La granularité de l'index donne l'exactitude avec laquelle on souhaite localiser une information. Elle peut aller du document entier à la phrase ou à la position du mot et elle définit aussi l'unité d'information qu'il sera possible de retourner aux utilisateurs.

Classiquement en RI, la tâche d'indexation d'un document textuel est basée sur la loi de *Zipf* [164] et sur la conjecture de *Luhn* [97]. La loi de *Zipf* stipule qu'en dressant une table de l'ensemble des mots différents d'un texte quelconque, classés par ordre de fréquence décroissante, la fréquence d'un mot est inversement proportionnelle à son rang dans la liste. Autrement dit, le produit de la fréquence de n'importe quel mot par son rang est constant. Ainsi, l'analyse statistique des documents textuels en anglais montre que les mots les 20% les plus fréquents représentent 70% du vocabulaire des documents écrits (*Salton* [127]). *Luhn* a utilisé cette loi pour déduire une relation entre la fréquence des termes et leur pouvoir discriminant (cf. figure 2.4). L'utilisation de la loi de *Zipf* et de la conjecture de *Luhn* permettent ainsi la mise en place d'un processus d'indexation généralement en trois phases :

Extraction : Elle consiste à collecter les mots qui ont un intérêt quant à la description sémantique du document. La collecte peut être manuelle (coûteux, subjectif et pouvant donner

FIG. 2.4 – Pouvoir d'expression des mots : conjecture de *Luhn*



lieu à des problèmes de consistance) et/ou (semi)automatique. De nombreux travaux en RI se sont penchés sur la question de l'efficacité de l'une ou l'autre des techniques d'indexation (*Witten et al.* [162] donnent un éventail de ces travaux).

Sélection : Une fois les termes les plus intéressants sélectionnés, des transformations sur les informations peuvent avoir lieu : élimination de mots vides (mots très fréquents et/ou sans portée sémantique forte tels que les mots de liaison). Une phase de lemmatisation est souvent mise en œuvre qui consiste pour chaque mot à ôter les suffixes et/ou préfixes de mots pour n'en garder que la racine, convertissant en une forme neutre où le genre et le temps (de conjugaison) sont enlevés. Des changements de casse sont également adoptés dans certains cas.

Pondération des termes d'indexation : La pondération des termes vise à établir des statistiques sur la collection qui seront utilisés pour déterminer les documents pertinents. Les techniques de pondération des termes s'appuient couramment sur la fréquence des termes dans un document (*tf* pour *term frequency*, mesure représentant l'importance locale d'un terme - fréquence relative) et de fréquence de ces termes dans l'ensemble des documents de la collection (on utilise l'inverse de cette mesure, simulant l'importance globale d'un terme, appelé *idf* pour *inverse document frequency* - fréquence absolue). Le tableau 2.6 de la section 2.5.1.0 présentera quelques fonctions *tf* et *idf* utilisées pour calculer ces deux facteurs dans certains modèles de RI.

Les trois techniques principales pour la mise en place d'index sont l'utilisation des fichiers inversés (cas le plus courant), les *signature files* et les *bitmaps* ; *Witten et al.* ([162], chapitre 3) discutent en détail de chacune de ces techniques avec ses avantages et inconvénients et des conditions dans lesquelles une technique est préférable à l'autre.

Face à la croissance en volume des collections, le coût de stockage et de maintien des index

augmente. Des techniques pour réduire l'espace de stockage des index et pour faciliter leur mise à jour ont été proposées et portent sur la factorisation conceptuelle de l'information et sur la compression physique des données contenues dans l'index.

2.4.2 Point de vue de l'espace de stockage

La taille d'un fichier inversé peut être réduite considérablement en utilisant les techniques de compression. Un panel des techniques de compression de textes est étudiée par*Baeza-Yates et Ribeiro-Neto* ([10], chapitre 2). L'adaptation de ces techniques au cas des index de type fichiers inversés part de l'observation que chaque liste inversée contenant un terme et l'ensemble des documents au sein desquels il apparaît peut être transformée sans perte d'informations comme une séquence d'entiers croissants, en stockant le premier document suivi d'une liste de *d-gaps* donnant les différences entre numéros de documents ; ceci permet de coder les listes inversés en utilisant moins d'espace. Divers travaux se sont interessées à l'utilisation des *d-gaps* (une liste de ces travaux est fournie dans [10], page 116). Divers travaux proposent plus généralement des solutions portant sur la compression physique des index ([110], [141], [72]).
Un *signature file* est une méthode probabiliste pour indexer du texte. A chaque document est associé un descripteur (ou signature), une suite de bits qui caractérise en un certain sens le contenu du document puisqu'elle est construite à partir des termes du document.
Un *bitmap* est une structure d'indexation simple : pour chaque terme, un *bitvector* est sauvegardé, chaque bit correspondant à un document. Le bit est positionné à 1 si le terme apparaît dans le document et à 0 sinon. Les *bitmaps* sont rapides et faciles d'utilisation mais ils consomment énormément d'espace de stockage. Ils sont particulièrement adaptés pour les requêtes booléennes.
Les fichiers inversés sont presque toujours supérieurs aux *signature files* et aux *bitmaps* dans la pratique, en terme d'espace de stockage et de temps nécessaire pour répondre aux requêtes ([162], page 130, chapitre 3) mais des combinaisons de ces trois techniques ont aussi été proposées.
L'utilisation de *stop list* (ou autrement dit l'omission des mots-vides dans l'index) réduit la taille de l'index mais pour *Witten et al.* [162], il est possible d'obtenir un gain d'espace même en gardant tous les mots dans l'index, quite à ne pas tenir compte des mots-vides lors du traitement des requêtes. Les phases de lemmatisation et de changement de casse permettent également une réduction de la taille des index (pour les trois techniques d'index citées précédemment). Pour les fichiers inversés, il y a en effet moins de listes inversées à stocker et celles stockées sont moins longues, rendant leur stockage moins coûteux. Pour les *signature files* ceci réduit le nombre de termes distincts, réduisant ainsi la largeur de la signature et pour les *bitmaps* le nombre de termes à indexer est également reduit. La liste des mots-vides peut être étendue à des termes fréquem-

ment utilisés pour réduire la taille de l'index. Ceci implique toutefois que les futurs utilisateurs sont privés de l'inclusion de ces termes dans leurs requêtes, ce qui peut être frustant (voir [162] pour plus de détails sur les problèmes que peuvent engendrer ce type de mots vides).
La représentation conceptuelle de l'information implique la réduction de l'espace de représentation d'une collection car les documents sont représentés à l'aide de concepts agrégés plutôt que des unités d'informations fines comme les termes [43], [21]. Les solutions apportées ont toutefois été validées sur des collections nettement moins volumineuses que celles produites actuellement [28].
Pour ce qui concerne les données multimédias, la *malédiction de la dimension* fait référence aux difficultés de gestion et de traitement de données qui apparaissent dans des espaces de grande dimension. Elle est liée d'une part, au très grand nombre de descripteurs de bas niveau que l'on peut extraire des images bandes audios et vidéos et, d'autre part, aux dimensions importantes de la plupart des descripteurs (histogrammes, transformées de Fourrier,...). Il est en effet difficile d'étendre dans ces espaces les techniques que l'on a dans les espaces à deux ou trois dimensions. Les tailles des descripteurs posent des problèmes de réduction de la dimensionnalité qui ont été explorés par différentes techniques comme l'analyse en composantes principales, les SOM pour *Self Organising Maps*, les SVM pour *Support Vector Machine*). Des travaux comme la thèse de *Berrani* [20] (chapitre 2) présente ces techniques .

2.4.3 Point de vue du temps

Le temps d'indexation moyen de la collection augmente de manière très significative en fonction de la taille des collections [156], [70].
Nous constatons comme *Chevallet et al.* [28] que les solutions proposées au problème de la représentation de grands volumes d'informations sont parcellaires : soit elles abordent un niveau conceptuel (peu de travaux), soit elles résultent d'une technique liée à des spécifications d'ordre physique (espace de stockage principalement, nombre d'entrée / sortie, etc.). Il serait intéressant d'aller vers des solutions basées sur la conceptualisation des index, l'objectif visé étant d'intégrer dans un même modèle les contraintes des différents niveaux d'abstraction afin d'en déduire des règles de conception de granules d'information.
Une fois les documents représentés au sein de l'index, les demandes des utilisateurs peuvent être traitées de façon plus aisée.

2.5 Passage à l'échelle pour l'Interrogation

L'interrogation est la phase du processus de RI qui se déclenche à chaque nouvelle demande de l'utilisateur. Trois étapes principales sont présentes dans cette phase : tout d'abord la formulation du besoin d'information de l'utilisateur appelée aussi requête, ensuite la recherche dans le corpus d'informations et enfin la présentation des résultats (cette dernière étape peut être vue comme une phase à part entière du processus de RI et c'est pourquoi nous la traitons à part dans la section 2.6). Le cadre théorique de l'interrogation est donc composé d'un *modèle de requête* et d'une *fonction d'appariement*.

- Le modèle de requête décrit comment le besoin d'information de l'utilisateur doit être formulé pour être compris et traité par le SRI. Des travaux se sont penchés sur la façon d'aider l'utilisateur dans la formulation de son besoin d'information. Les modèles de RI prenant en compte le retour de pertinence permettent également à l'utilisateur d'affiner son besoin d'information au cours de la recherche.
- La fonction d'appariement identifie les documents correspondant à la requête de l'utilisateur ou documents pertinents (pertinence système). Pour ce faire, elle compare la représentation de la requête à celles des documents et elle fournit une liste de documents-résultats. Cette fonction qui est fortement corrélée aux représentations (des documents et de la requête) est une des principales caractéristiques qui différencie les modèles de RI.

2.5.1 Fonctions d'appariement

Il existe plusieurs modèles de RI ayant chacun sa technique de représentation de documents et de requête et sa façon de les apparier. Les modèles traditionnels ont été décrits de façon détaillée ([123], chap 4) et ([10], chap 2). Une description brève de l'espace de représentation (document, requête) associée à quelques modèles classiques de RI est donnée dans les paragraphes suivants.

Modèles ensemblistes

Le modèle booléen est un modèle classique en RI qui fait partie des modèles ensemblistes de RI ; dans ce modèle, chaque document est représenté par l'ensemble des termes qui le composent et la requête de l'utilisateur est formulée à l'aide d'une expression booléenne (des termes et des sous-expressions connectés par des opérateurs booléens). C'est un modèle dit exact car il ne renvoie que les documents répondant de façon exacte à la requête. Il sépare alors le corpus de documents en deux groupes (documents vérifiant la requête ou non) et ne permet donc pas d'ordonner les documents à retourner à l'utilisateur. Pour graduer l'appartenance au groupe des

TAB. 2.5 – Score obtenu pour la requête booléenne $(t_1\, Ou\, (\, Non(t_2)\, Et\, t_3\,))$ en fonction du contenu du document D

$t_1 \in D$	$t_2 \in D$	$t_3 \in D$	score de pertinence de D
1	x	x	1
0	1	x	0
0	0	0	0
0	0	1	1

documents pertinents, plusieurs modèles basés sur la théorie des sous-ensembles flous ont été développés [108]. Le modèle booléen a également été étendu pour traiter les poids des termes et permettre d'ordonner les documents à retourner [128]. Des extensions du modèle booléen ont aussi été proposées pour prendre en compte, entre autres, la proximité entre les occurrences des termes de la requête dans le document en introduisant un opérateur *NEAR* [129]. La table 2.5 donne un exemple d'appariement entre un document et une requête dans le modèle booléen de base.

Modèles algébriques

Le modèle vectoriel fait quant à lui partie des modèles algébriques. Les requêtes et les documents sont représentés par des vecteurs dont les composantes sont les poids pour les termes du vocabulaire. Classiquement, le poids $w(d,t)$ du terme t dans le document d dépend de la fréquence du terme (tf) dans ce document et de la fréquence documentaire de ce terme, c'est-à-dire du nombre de documents où le terme t apparaît. En pratique, plusieurs fonctions pour le calcul de la fréquence du terme et pour l'inverse de la fréquence documentaire (idf) ont été utilisées. Ces deux données sont à la base des modèles communément appelés *tf.idf*, c'est-à-dire des modèles où le score de pertinence d'un document à une requête est d'autant plus élevé qu'il contient de nombreuses fois les termes de la requête de l'utilisateur (facteur *tf, term frequency*), et que ces termes discriminent les documents entre eux (facteur *idf, inverse document frequency*). La valeur de similarité entre le vecteur représentant un document et le vecteur représentant une requête donne la pertinence de ce document par rapport à la requête ; cette similarité est calculée par un produit scalaire ou (le plus souvent) par la mesure du cosinus de l'angle entre les deux vecteurs [129]. La table 2.6 donne quelques unes des fonctions implémentées dans certains modèles pour représenter ces deux facteurs.

TAB. 2.6 – Exemples de fonctions *tf* et *idf* [15]. $f(d, t)$ est le nombre d'occurrences du terme t dans le document d, df(t) est le nombre de documents au sein desquels le terme t apparaît, $\mid D \mid$ est le nombre total de documents.

Fonctions *tf*	image	Fonctions *idf*	image
$tf(d, t) = f(d, t)$	$[0, 1]$	$idf(t) = \frac{\log(\mid D \mid}{df(t)}$	$[0, \log(\mid D \mid)]$
$tf(d, t) = \frac{f(d,t)}{max_{t'}(f(d,t'))}$	$[0, 1]$	$idf(t) = \frac{1}{df(t)}$	$[\frac{1}{\mid D \mid}, 1]$
$tf(d, t) = 1 + \log(f(d, t))$	$[0, +\infty]$	$idf(t) = \log(\frac{\mid D \mid - df(t)}{df(t)})$	$[-\infty, \log(\frac{\mid D \mid - 1}{\mid D \mid})]$

Modèles probabilistes

L'approche probabiliste en RI a été initiée par *Maron et Kuhns* en 1960 [102]. Depuis, elle est mise en œuvre par diverses techniques dans les travaux de *Maron et Cooper*, ceux de *van Rijsbergen* ou encore de *Croft et Turtle* et de *Robertson*. Ces modèles traitent la pertinence d'un document par rapport à une requête à travers le principe de classement probabiliste (*Probability Ranking Principle*) expliqué par *Robertson* [124] et qui stipule que la meilleure recherche se fait dans le cas où les documents sont retournés par ordre décroissant sur leur probabilité de pertinence par rapport à la requête[23]. *Fuhr* [53] dresse l'éventail des différents modèles probabilistes. Il existe d'autres modèles de RI : les modèles alternatifs des trois grandes familles décrites précédemment sont présentés par *Baeza-Yates et Ribeiro-Neto* [10] ; les travaux de *Hawking et Thistlewaite* [68], de *Clarke* [30], de *Rasolofo* [120] et de *Mercier* [107] proposent des modèles basés sur la proximité entre termes de la requête.

Dans le cadre de la recherche dans les documents structurés, les différents types de modèles sont étendus de diverses façons pour tenir compte de la structure ou, le plus souvent, des impacts de la présence d'une structure sur les caractéristiques des unités d'information. On ajoute ainsi des paramètres supplémentaires pour ajuster les formules classiques :

- le nombre de sous-éléments d'un élément est important (*Fuller et al.* [54]),
- ainsi que son type (*Fuller et al.* [54], *Dopichaj* [42]),
- la fréquence de ce type d'élément dans la collection (*Schlieder* [140], *Grabs et Schek* [59]),
- l'importance d'un terme dans les autres éléments du même type (*Grabs et Schek* [59])
- ou tout simplement des index séparés par types d'éléments (*Mass et Mandelbrod* [103]).

L'accroissement des volumes des collections a un impact évident sur l'étape de traitement de requêtes tant sur l'*efficience* (*i.e.* la capacité de rendement, la performance) que sur l'*efficacité*

[23] Optimum retrieval is achieved when documents are ranked according to decreasing values of the probability of relevance (with respect to the current query) (extrait de [124])

(*i.e.* le fait de produire l'effet attendu).

2.5.2 Efficience et efficacité face à la croissance

Rôle de la collection dans l'attribution de score aux documents

Pour fournir une liste de résultats ordonnés en réponse à une requête, les SRI attribuent à tout document de la collection interrogée un score de pertinence (*Retrieval Status Value, RSV*). Les documents sont ordonnés suivant ce score de pertinence avant d'être retournés à l'utilisateur (sauf dans le cas des modèles exacts comme le modèle booléen de base). Le score de pertinence est donc un concept central dans la phase d'interrogation et de retour de résultats. Le rôle que joue la collection (et donc la taille de la collection) sur l'attribution de ce score de pertinence peut avoir une influence sur les performances du SRI dans les collections volumineuses.
Les travaux de *Hawking et Robertson* [67] sur le rôle de la collection de documents dans l'attribution de score de pertinence ont permis de classer les modèles de RI en trois groupes :

- *Invariance déterministe* : les modèles de RI pour lesquels le score d'un document reste le même quelle que soit la collection au sein de laquelle est pris ce document. L'attribution d'un score à un document ne dépend que de la requête et du document et pas de la collection. C'est le cas du modèle booléen (de base) et des modèles de RI qui calculent le score d'un document par rapport à une requête en se basant sur la proximité entre les termes de la requête au sein du document (par exemple les modèles de *Hawking et Thistlewaite* [68], de *Clarke et al.* [30] et de *Mercier et Beigbeder* [107]).
- *Invariance statistique* : il existe des modèles pour lesquels la formulation de la requête [24] est indépendante de la collection mais le calcul du score d'un document a un paramètre prenant en compte la collection[25]. Toutefois, l'algorithme d'attribution de score se comporte statistiquement de façon identique quelle que soit la collection. C'est par exemple le cas des modèles utilisant l'*idf*. Un même document n'aura pas forcément le même *idf* suivant la collection au sein de laquelle il se trouve mais la formulation de la requête n'est pas influencée par la collection.
- Les modèles de RI dont la formulation de la requête varient d'une collection à une autre pour une même requête. C'est le cas des systèmes à deux passes ou utilisant le *retour de pertinence aveugle* (*blind feedback systems*). Les premiers résultats d'une requête sont utilisés pour étendre la requête pour une seconde passe ; ainsi, la formulation finale de

[24] *search formulation* d'après les auteurs

[25] the *search formulation* is essentially independent of the collection (some form of query tokenisation for e.g.) but the weights assigned to terms are collection-dependent.

la requête (les termes tout autant que les poids) dépend fortement de la collection. La collection affecte fondamentalement le processus d'attribution des scores.

Prenant en compte ces différents cas, les auteurs étudient l'impact de la taille de collection sur les performances en RI en utilisant quelques métriques classiques. Ces études sont présentées en détails dans le chapitre 3. Nous avons utilisé des modèles de RI satisfaisant les deux premiers cas dans notre étude et d'après nos résultats, le rôle de la collection dans l'attribution des scores de pertinence agit sur le passage à l'échelle des modèles ; nous présentons ces résultats dans le chapitre 4.

Efficience et efficacité

Contrairement à la phase d'indexation où c'est le rendement qui est le plus important, durant l'interrogation c'est le temps de réponse qui importe le plus. L'impact de l'accroissement des volumes des collections est donc lié à la complexité en temps de la fonction d'appariement. En effet, il peut être intéressant d'avoir un temps de réponse qui n'augmente pas avec la taille de l'espace de recherche. Le temps moyen de traitement des requêtes augmente pourtant de manière très significative en fonction de la taille des collections d'après [156, 70] (cités par [28]). La piste *Terabyte* de TREC initiée depuis 2004 propose une tâche portant sur l'efficience des systèmes de RI mais s'intéresse plutôt au temps d'exécution des requêtes. Ainsi, les temps d'exécution des requêtes sont déterminés pour les différentes SRI participants. Les temps d'exécution des huit meilleurs systèmes (classés suivant la précision après vingt documents retournés) varient entre 1 201 et 219 354 secondes[26]. Toutefois, il n'est pas possible de comparer directement les performances des systèmes en temps puisqu'ils utilisent des matériels très différents (ordinateurs personnels, supercalculateurs,...). Des tentatives de normalisation du temps d'exécution des requêtes par le coût estimé du matériel utilisé ou par le nombre d'unités centrales utilisées ont été proposées pour *TREC 2005*, sans permettre toutefois de faire des comparaisons effectives entre systèmes [31].

L'appariement consiste tout d'abord à évaluer une requête face à l'ensemble des documents et ensuite à effectuer un classement de ces documents suivant le score de pertinence. Ce classement peut devenir problématique si on ne limite pas très rapidement le sous-ensemble de documents sur lequel va porter le tri. Certains modèles évitent le tri explicite et exploitent plutôt des heuristiques réputés fiables [28]. Par exemple, certains modèles vectoriels effectuent le classement décroissant des documents vis-à-vis de chaque terme pour en dériver un classement décroissant

[26] C'est le temps moyen mis pour retourner les 20 premiers documents. Les systèmes utilisaient un nombre variable d'unités centrales avec des caractéristiques variables également

des documents vis-à-vis de la conjonction des termes). De façon générale, il convient de transférer aussi largement que possible la complexité en temps dans la phase d'indexation.

Réduire la taille de l'ensemble sur lequel s'effectue la recherche en identifiant des sous-collections au sein de l'ensemble des données permet aussi de réduire le temps de traitement de requête. La difficulté est alors de déterminer les bases sur lesquelles la segmentation devra être effectuée (profil des utilisateurs [112], méta-données le besoin en information de l'utilisateur).

L'impact de l'accroissement des corpus de documents sur l'efficacité est lié à la quantité de documents qu'il faut classifier selon leur pertinence vis-à-vis d'une requête. De façon générale, les travaux en RI se sont peu intéressés à l'impact de la taille des collections sur l'efficacité. D'après [28], il est intuitivement plus *facile* d'identifier les documents pertinents dans une collection comportant quelques centaines de documents plutôt que dans une collection comprenant des milliards de documents (le risque d'erreur de classification est statistiquement plus faible dans le premier cas), et les résultats obtenus dans les campagnes *TREC* confirment ce point [70]. *Salton et McGill* [129] affirmaient dans les années 1980 que la précision devrait diminuer quand la taille de la collection augmente parce que le nombre de documents pertinents retournés ne semble pas devoir augmenter en proportion avec la taille de la collection[27]. Il n'est pas aisé de vérifier cette prédiction sur les collections de test actuelles puisque seule une partie de leurs documents pertinents est détectée ; un bémol est à introduire concernant cette prédiction puisque l'accroissement de collections signifie aussi l'ajout de nouveaux documents éventuellement pertinents, ce qui peut produire des résultats surprenants comme ceux obtenus dans certaines études [69, 67] et dans des travaux que nous avons menés et que nous expliquons au chapitre 4 concernant l'amélioration de la haute précision quand la taille de collection augmente.

De façon pratique, l'augmentation des volumes à traiter nécessite de procéder à une parallélisation massive des traitements, à la manière des moteurs de recherche sur le Web qui exploitent à ce jour une grappe de plusieurs milliers de machines (serveurs de stockage et/ou de calcul). Il faut noter toutefois que le monde académique est limité face à ce type de pratique car un laboratoire investira difficilement dans l'achat de nombreuses grappes de stockage et/ou de calcul, comme l'ont remarqué *Chevallet et al.* [28].

La requête posée et l'appariement effectué, les systèmes de RI sont prêts à fournir les références vers les unités d'informations qu'ils ont trouvé les plus intéressantes. Le passage à l'échelle n'est pas sans effet sur cette étape importante.

[27] *the number of retrieved relevant documents is not likely to increase in proportion to the size of the collection*

2.6 Retour de résultats et passage à l'échelle

Lorsqu'on envisage l'accès par recherche à de grandes masses d'informations, il faut également traiter de la présentation des résultats. Nous nous intéressons dans cette section à savoir comment le passage à l'échelle influence les modes de présentation des données et particulièrement le retour des résultats en RI. De façon générale, les travaux sur la visualisation de grandes masses de données ont été réalisés principalement dans le domaine des interfaces homme-machine ([73, 113]). Quelques travaux spécifiques à la visualisation en RI ont également été menés ([27, 116, 71]).

2.6.1 Techniques de visualisation d'information

Pour rendre agréable la communication entre les utilisateurs qui recherchent de l'information et les systèmes de RI, il est nécessaire de mettre en place des interfaces entre eux. En effet, comme mentionné précédemment, le besoin d'information de l'utilisateur est généralement peu clair pour lui même, et il varie au cours des interactions avec le système de RI ou simplement au cours du temps. Le but d'une interface sera donc de favoriser l'étape de formulation du besoin d'information (expression par une requête, choix des sources au sein desquelles va s'effectuer la recherche), de permettre une bonne compréhension des résultats retournés par le système de RI et éventuellement de réaliser un suivi pour une recherche sur un sujet donné en créant par exemple des profils d'utilisateurs, ou encore en prenant en compte l'avis des utilisateurs sur les résultats retournés.

La subjectivité et la complexité de l'humain (de son raisonnement) rendent peu aisée la mise en place d'interfaces permettant de venir à bout des buts ci-dessus cités de façon efficace, surtout que la culture, l'âge et d'autres critères comme le niveau de connaissance de l'utilisateur agissent sur sa satisfaction. Une interface intuitive pour un utilisateur averti semblera complexe pour un novice ; de même une interface fournissant plusieurs fonctions avancées de recherche sera utile pour un expert mais pour un non initié elle sera source de confusions. La mise en place d'interfaces en SI de façon générale et particulièrement dans le cas de la RI (où l'interaction avec le système de RI est primordiale) doit prendre en compte ce genre d'aspects. Elle doit également tenir compte autant que possible du champ d'étude auxquels va s'appliquer la RI (RI dans une base bibliographique, RI sur des documents textuels complets, thème sur lesquels portent les documents, etc).

L'interface joue donc un rôle prépondérant dans les phases d'interrogation, de visualisation des résultats de recherche. La rapidité sans cesse grandissante des processeurs d'images et de graphiques divers ainsi que la haute résolution en couleur des écrans actuels encouragent les progrès

en visualisation d'information par l'utilisation d'images (icônes, graphiques divers, etc). S'il est plus intéressant de représenter un visage humain par une image plutôt que par une description textuelle, il est moins évident d'utiliser des images pour des descriptions textuelles plus abstraites. Des techniques ont donc été mises au point pour permettre l'accès aisé aux résultats de recherche sous forme de documents textuels. Elles sont particulièrement intéressantes dans le cadre de collections volumineuses.

Pour *Baeza-Yates et Ribeiro-Neto* ([10], chapitre 10), les interfaces de RI doivent fournir aux utilisateurs un bon point de démarrage de leur recherche ; ces derniers étant connus pour leur tendance à débuter par des requêtes courtes qui sont ensuite modifiées ou raffinées en fonction des résultats retournés. Parmi ces interfaces pour le démarrage de la recherche, on peut citer des interfaces pour la sélection des sources au sein desquelles va s'effectuer la recherche, des interfaces fournissant un aperçu du contenu des différentes sources, des interfaces fournissant un exemple ou des *Wizards* à l'utilisateur.

Pour ce qui concerne la formulation de la requête elle même, les cinq styles primaires d'interaction homme-machine identifiés par *Schneiderman* (à savoir le langage de commande, le remplissage de formulaire, la sélection de menus, la manipulation directe et l'utilisation du langage naturel) ont été utilisés dans différentes interfaces de formulation de requêtes dans les systèmes de RI. La manipulation directe implique notamment des approches graphiques pour les interfaces de formulation de requêtes.

Le retour de résultats de recherche se fait en général sous forme d'une liste ordonnée de références par exemple des liens hyper-textes vers les documents pour les moteurs de recherche sur le Web, l'ordre des références donne la probabilité de pertinence des documents. Des efforts pour placer les résultats de recherche dans leur contexte initial au sein de la collection ont vu le jour ; les propositions s'articulent autour de la mise en exergue du contexte d'apparition des termes de la requête ou alors l'affichage de liens qui existent entre documents de la collection.

Parmi ces propositions, celles utilisant les *KWIC (KeyWord-In-Contexte)* montrent les phrases qui résument la façon dont les termes de la requête sont utilisés dans le document. Les *TileBars*[28] quant à eux permettent d'afficher pour chaque document une barre graphique montrant son degré de correspondance à chaque *facette* d'une requête. Des systèmes permettant de regrouper les documents en fonction du sous-ensemble des termes de la requête qui y apparaissent existent également (*InfoCrystal*, *VIBE*). Le système *SuperBook* utilise la structure d'un document de grande taille pour afficher les termes de la requête dans leur contexte d'apparition. L'organisation des documents en catégories selon leur contenu est également utilisée pour la visualisation

[28] traduction littérale : barre de tuile

des résultats (système *CORE* ou encore le système *DynaCat*). La description graphique de liens hyper-textes entre documents est une technique de visualisation de résultats mais aussi une façon de guider la recherche elle-même (système *Mapuccino*). L'organisation des documents sur des critères comme l'auteur ou la date est aussi proposée dans des systèmes comme le système *Envision*.
Enfin, des interfaces pour permettre la reformulation de requête par interaction avec le système de RI existent également. Ce type d'interface fournit à l'utilisateur les moyens d'attribuer aux documents une pertinence (binaire ou multivaluée), ce qui permet de générer une nouvelle requête et de relancer la recherche (voir [10], chapitre 10).

2.6.2 Visualisation en Recherche d'Information

Les solutions pour la visualisation des résultats sur les écrans visent à optimiser la surface affichée pour augmenter l'acuité de la perception. Cette optimisation est dépendante du média ; par exemple la perception globale d'une image et d'une mosaïque d'images est plus efficace que son contenu textuel [28]. Pour ce qui concerne les documents textuels et les images, actuellement les résultats (sur le Web) sont présentés généralement en une dimension : liste de documents (ou de liens vers des documents) classés par ordre de pertinence système. Certains moteurs de recherche proposent une deuxième dimension en créant une sous-organisation des réponses en les regroupant par leur site d'origine. Des métaphores sont aussi utilisées comme la loupe pour examiner une partie des réponses tout en conservant une perception globale. L'introduction d'une troisième dimension de visualisation peut être effectuée avec par exemple la métaphore 3D d'une bibliothèque avec les résultats dans des rayonnages visuels. L'utilité et l'efficacité réelle de ces propositions est difficile à apprécier, mais elles sont certainement adaptées à certains types de recherche et/ou d'utilisateurs. Certaines techniques de visualisation peuvent avoir une influence sur le modèle de RI : c'est le cas du modèle ostensif [26]. Il prend en compte la navigation de l'utilisateur parmi les résultats (donc son comportement) comme expression de son besoin. La navigation est donc vue comme un élément (complexe) de la rétro-action entre l'utilisateur et le système. Les possibilités offertes à l'utilisateur pour exprimer son besoin d'information devront également de plus en plus tenir compte des évolutions technologiques et aller vers une communication multimodale par la voix, l'image et la gestuelle.
Il y a peu d'études qui ont porté sur des applications de visualisation et navigation dans de grandes quantités de données en RI (les études existantes sont dédiées à des cas particuliers comme la visualisation d'images médicales, de la structure du Web, etc.). Les systèmes disponibles comme *Kartoo* n'ont pas été évalués. L'influence du passage à l'échelle porte sur le nombre de

réponses potentielles à une requête et pose donc le problème de la navigation parmi ces réponses. *Chevallet et al.* [28] proposent de permettre une perception du contenu des corpus pour aider l'utilisateur à mieux établir son propre jugement de pertinence.

2.7 Synthèse

L'accroissement de la quantité d'information numérique a été examiné dans ce chapitre au travers de ses causes et de quelques études quantitatives. Les moteurs de recherche (commerciaux) font face à cette croissance et réagissent chacun à leur façon. Il est en effet à noter que, certaines sinon toutes les phases du processus de RI se trouvent questionnées par l'accroissement continu et rapide du nombre de documents. Les enjeux à prendre en compte face au passage à l'échelle pour chacune de ces phases ont été examinés. Les principales directions concernant chacune des phases ont également été mises en exergue. Elles portent principalement sur la reduction de l'espace de recherche (par prise en compte de l'utilisateur pour situer son profil par rapport au contenu de collections, par une organisation et une représentation de collections volumineuses intégrant différents niveaux d'abstraction de l'information), l'adaptation des techniques de visualisation de grandes quantités d'informations à la tâche de RI.
D'autres facteurs relatifs aux modèles de RI, influent sur les solutions connues et vont nécessiter de nouveaux apports. De nouvelles directions pour le passage à l'échelle sont liées à l'hétérogénéité des corpus documentaires et à l'apparition de nouveaux médias. En effet, l'hétérogénéité est plus importante dans les collections volumineuses et les descriptions statiques ne sont plus discriminantes. *Chevallet et al.* [28] proposent de revoir par exemple la conjecture de *Luhn* dans le cas de corpus de très grandes dimensions. Il semble de même que l'hypothèse qui prévaut à l'usage de la fréquence documentaire inverse ne soit pas valide dans les grandes collections car l'influence du facteur *idf* va en diminuant avec la taille des collections [15].
L'évaluation est le moyen privilégié pour déterminer l'évolution du domaine et permettre la comparaison de SRI. Elle est également affectée par le passage à l'échelle et des réflexions sont menées pour élaborer des méthodologies d'évaluation des performances de recherche dans les collections volumineuses. Nos travaux (chapitre 5) ont porté sur cet aspect et c'est pourquoi l'évaluation en RI est le centre d'intérêt du chapitre suivant.

Chapitre 3

Evaluation en RI et Passage à l'échelle

Sommaire

Une des phases clé de la mise en place de tout système d'information est la phase d'évaluation. Son but est de déterminer si les fonctionnalités attendues du système sont mises en œuvre et si elles respectent les contraintes de l'environnement de travail dans lequel le système sera utilisé. C'est l'évaluation fonctionnelle du système lui-même. Il existe un autre type d'évaluation des systèmes de RI qui s'intéresse non pas au système de RI lui même, mais qui vise plutôt à établir la qualité de recherche réalisée par le système. Pour mener à bien cette dernière évaluation, des

techniques d'évaluation et des métriques ont été proposées dans divers travaux de RI, analysant la qualité des résultats de recherche tantôt du point de vue du système de RI, tantôt du point de vue de l'utilisateur du système de RI. Au centre de toutes ces métriques et au centre de l'évaluation en RI se trouve la notion de pertinence, qui détermine le degré auquel une unité d'information donnée correspond (apporte satisfaction) à un besoin d'information donné. La pertinence est binaire ou multivaluée selon les études. nous nous intéressons aux métriques principales existant dans le domaine de la RI et nous analysons cette notion complexe et multidimensionnel qu'est la pertinence. Nous nous intéressons ensuite à déterminer les aspects sur lesquels il faudra focaliser l'évaluation lorsque l'on travaille dans des environnements à quantité de documents de plus en plus volumineuse.

3.1 Introduction

L'évaluation est une phase clé dans le processus de mise en œuvre de tout système d'informations. Pour la mener à bien, il est nécessaire d'établir de façon claire les fonctionnalités attendues du système, les contraintes à respecter par le système (qui dépendent de l'environnement de travail au sein duquel il sera utilisé). Pour ce qui est des fonctionnalités, leur choix et leur orientation se fait de façon générale avec l'avis des futurs utilisateurs du système. Des méthodes d'analyse et de conception de systèmes d'information existent pour assurer la prise en compte des besoins réels des utilisateurs tout au long du processus de mise en œuvre. Une fois cette mise en œuvre réalisée, il est presque naturel de faire le bilan et d'établir la façon dont le système atteint les objectifs qui lui étaient assignés. Pour ce qui concerne les contraintes liées à l'environnement au sein duquel le système sera utilisé, elles peuvent concerner divers aspects comme le temps de réponse ou l'interface du système (par exemple simplicité d'utilisation, intuitivité et convivialité). Les critères d'évaluation des systèmes d'informations ont fait l'objet de nombreuses études. Divers critères permettent d'évaluer la façon dont les systèmes d'informations mettent en œuvre les différentes fonctionnalités. La plupart des systèmes d'informations peuvent être évaluées de façon binaire en ce qui concerne des fonctionnalités liées au traitement des informations (exemple : établir la liste des employés affectés à une agence, calcul du salaire moyen des employés de telle catégorie, etc.). Soit le système met en œuvre la fonctionnalité de façon correcte, soit il ne le fait pas. En recherche d'information, en plus de ce type d'évaluation, il est nécessaire d'établir la qualité des résultats fournis par le système de RI. La qualité des résultats est délicate parce que subjective ; elle peut être observée de deux points de vue qui ne sont pas toujours concordants : le point de vue du système de RI et le point de vue de l'utilisateur.

- le point de vue de l'utilisateur est lié directement à son besoin d'information. Ce besoin est vague. En effet, deux utilisateurs peuvent faire deux descriptions différentes d'un même besoin d'information et la description d'un besoin d'information donné pour un même utilisateur est tout aussi variable en fonction de l'environnement au sein duquel il se trouve ou entre deux instants (*Schamber* [138]). De même, une même formulation peut correspondre à deux besoins d'informations réels différents. Les travaux de *Mizzaro* [109] montrent plusieurs étapes du besoin réel d'un utilisateur à la requête finale soumise à un système de RI. Ces étapes entraînent la perte d'information (ou la transformation d'information en tout cas) entre la réalité du besoin d'information et son expression. En fonction du langage de requête utilisé, cette perte est plus ou moins importante. Les travaux portant sur l'utilisation du langage naturel permettent un rapprochement entre le besoin réel et sa formulation en requête. Toutefois, les phénomènes linguistiques de variations morphologiques, lexicales, sémantiques et syntaxiques des mots sont des freins auxquels fera encore face cette approche. La prise en compte du contexte est une piste qui permet également de rapprocher le besoin d'information de sa formulation.
- le point de vue du système est lié à la fonction d'appariement entre la représentation que le système fait de la requête (qui est elle même déjà une représentation partielle du besoin d'information réel) et la représentation que le système fait du document. Le document est l'œuvre d'un humain et est donc également empreint de subjectivité et de versatilité. Comme nous l'avons vu dans le chapitre 2, divers modèles de RI implémentent diverses façons de représenter documents et requêtes et diverses manières de les apparier.

La tâche des systèmes de RI est de concilier ces deux points de vue ; les constats précédents nous montrent que c'est une tâche qui n'est point aisée, mais aussi que face à un besoin d'information, l'ensemble des bonnes réponses n'est pas unique (puisqu'il dépend pratiquement de l'utilisateur) ; dans l'absolu, il n'y aura donc même pas de « bons »résultats. La mise en correspondance des deux points de vue mentionnés ci-dessus se fait sous le concept de *pertinence*. Les constats précédents nous introduisent à la complexité de ce concept, qui induit des processus cognitifs et psychologiques. C'est le point principal que nous analysons dans cette partie. L'évaluation en RI n'a pas toujours tenu compte des tailles de collections. Il nous semble toutefois que dans les environnements volumineux, la problématique de l'évaluation ne se pose pas en des termes identiques à ceux d'environnements de taille réduite. La masse d'information étant plus importante (et plus hétérogène), le problème de famine d'information n'existe plus mais est remplacé par celui de surabondance d'information ; il est nécessaire de repenser ce qui doit être évalué dans les (résultats de recherche des) systèmes de recherche d'information.

Dans la section suivante, après avoir présenté les méthodologies d'évaluation des systèmes de RI, nous revenons sur la pertinence comme notion binaire ou multivaluée. Ces deux « classes »de pertinence ont donné lieu à des métriques d'évaluation différentes. Dans la section 3.4, notre intérêt porte sur les métriques d'évaluation à pertinence binaire et à leur utilisation pour évaluer la façon dont les systèmes de RI passent à l'échelle. Dans la section 3.5, les principales (et récentes) métriques prenant en compte la pertinence multivaluée sont présentées et nous introduisons les travaux que nous avons menés pour évaluer le passage à l'échelle des systèmes de RI dans des environnements à pertinence multivaluée.

3.2 Evaluation de SRI

3.2.1 Tour d'horizon de l'évaluation

L'évaluation est comprise de façon générale comme un moyen de mesurer un événement ou une chose. Pour les systèmes d'information, l'évaluation peut concerner les caractéristiques techniques des systèmes et / ou les changements induits par le système au niveau des utilisateurs et des organisations. Dans tous les cas, pour parler d'évaluation, il est nécessaire de déterminer ce qui doit être évalué, des critères d'évaluation donnant les objectifs de l'évaluation, des métriques basées sur ces critères et une méthodologie d'évaluation (*Saracevic* [133]). Chacun de ces points est présenté pour ce qui est du domaine de la RI, pour lequel plusieurs classifications de l'évaluation ont été proposées.

Définition et classifications

Hernon et McClure (cité dans [84]) affirment que[29]

> l'évaluation est le processus par lequel sont identifiées des informations à propos d'un système pour déterminer la qualité du service et le degré auquel le service atteint les objectifs.

Saracevic[133] parle de juger la performance ou la valeur d'un système, d'un processus ou d'un produit. *Meadow* [105] distingue deux catégories d'évaluation en RI : l'évaluation de performances qui décrit ce qui se passe durant l'utilisation du système de RI (ce que nous avons nommé efficience) et l'évaluation des sorties qui décrit les résultats du système de RI (ce que nous avons nommé efficacité).

[29] *Evaluation is the process of identifying and collecting data about specific services, establishing criteria by which their success can be assessed, and determining both the quality of the service and the degree to which the service accomplishes stated goals.*

L'approche traditionnelle pour l'évaluation en RI est celle de l'utilisation des collections à la *Cranfield* telle que décrite précédemment en 2.3.1. Les nouveaux paradigmes de l'évaluation accordent une place plus importante à l'utilisateur, à son environnement de travail et à la prise en compte de ses aspects cognitifs. Ainsi, *Nilan et al.* [114] par exemple ont étudié les critères des utilisateurs pour la pertinence et ont trouvé plus d'une trentaine de critères parmi lesquels la *serendipity*, la couverture, la déduction logique, la facilité d'utilisation, des critères de croyance. *Harter* [64] suggère la conception d'une nouvelles techniques pour l'évaluation[30]. *Kagolovsk et Moehr* [84] donnent un éventail des travaux s'intéressant à ces nouveaux paradigmes.

Questions importantes en évaluation

La question de l'évaluation est liée à celle de l'organisation des informations, de la spécification du besoin d'information et du retour des résultats.

Ce qui est évalué : La difficulté (confusion) est souvent de choisir le niveau auquel va se situer l'évaluation. *Saracevic* [133] détermine deux classes pour l'évaluation, toutes deux nécessaires et complémentaires, découpées en trois niveaux chacune :

- évaluation orientée système (*engineering level*, le niveau des entrées, *processing level* (de nombreux travaux en RI s'y sont intéressés),
- évaluation orientée utilisateur : le niveau des sorties (auquel de nombreux travaux en RI se sont intéressés), le niveau utilisateur, et un niveau dit social. Peu de travaux ont porté sur l'ensemble de ces niveaux, chacun focalise en général sur un niveau donné. Les deux derniers sont de plus en plus étudiés.

Hersh[74] parle de deux sortes d'évaluation : macro-évaluation et micro-évaluation. *Lancaster et Warner* [91] définissent trois niveaux d'évaluation :

- l'efficacité du système et l'interaction de l'utilisateur avec le système,
- le *cost-effectiveness evaluation*,
- le *cost-benefit evaluation*.

Les critères et métriques d'évaluation dépendent de ce qu'on évalue. Dans le cas de l'évaluation concernant le système, des méthodes existent en système d'informations. Si l'évaluation concerne les interfaces, on utilisera par exemple les critères ergonomiques de *Bastien et Scapin* [13]. Dans le cas de l'évaluation des sorties du système (cas le plus courant en RI), la qualité des résultats fournis est déterminée par des métriques quantitatives comme la précision.

[30] Les auteurs parlent de *brand-new evaluation instruments*.

Méthodologie d'évaluation : Les travaux de *Tague-Sutcliffe* [150] posent un certain nombre de questions considérées des plus importantes pour l'évaluation en RI, parmi lesquelles le choix des juges pour établir les jugements de pertinence et la valeur d'une approche analytique par rapport à une approche expérimentale. *Harter et Hert* [65] posent également des questions liées à la prise en compte des utilisateurs dans l'évaluation et à l'évaluation de systèmes de RI partie par partie. Comme pour les critères d'évaluation, la méthodologie choisie dépendra de ce qui est évalué. Pour l'évaluation des interfaces par exemple, des méthodes requérant la participation directe des utilisateurs peuvent être utilisées : faire des tests utilisateurs, recueillir des données au travers de questionnaires [14]. Pour l'évaluation de la qualité des résultats, la méthodologie d'évaluation consiste à mettre en œuvre des collections avec jugements de pertinence, tel que c'est le cas pour les campagnes d'évaluation actuelles en RI. Les métriques de RI visent donc à évaluer les résultats de recherche pour la plupart et se basent sur la pertinence par rapport aux thèmes contenus dans les documents (*topicality*). Les métriques d'évaluation ont surtout porté sur l'évaluation de la qualité des résultats de recherche, dans un environnement où la RI se fait en mode batch. La prise en compte des utilisateurs en phase d'évaluation est de plus en plus courante, à travers des systèmes de RI interactive (piste *Interactive Information Retrieval* de *TREC*). Dans des contextes interactifs, des métriques d'évaluation spécifiques prenant en compte les interfaces sont nécessaires. Il faudra alors utiliser/adapter les travaux en Interaction Homme-Machine qui ont étudié les critères pour évaluer les interfaces par exemple et leur impact sur les tâches des utilisateurs (intuitivité et facilité d'utilisation, etc.).

Dans le monde de la recherche en RI, l'évaluation porte généralement sur la qualité des sorties rendues par les systèmes et le critère principal est la pertinence des documents par rapport à la requête. En ce qui concerne la méthodologie d'évaluation, elle s'appuie sur la construction de collections de test présentées dans le chapitre 2 et est fortement liée à la façon de les mettre en place. Dans la section suivante, nous expliquons la technique du *pooling* et donnons les travaux principaux s'y rapportant.

3.2.2 Construction de collection de test : le *pooling*

Méthodologie

Pour la construction de collections de test, le corpus de recherche est tout d'abord créé. Dans le cadre de *TREC*, suivant les années et les pistes de recherche auxquelles le corpus est dédié, il peut être constitué de pages Web, d'articles de journaux, etc. L'ensemble de besoins

FIG. 3.1 – Exemple de topic : le topic 460 de *TREC*

<top>
<num> Number : 460
<title> Who was Moses ?
<desc> Description : Find documents that discuss the biblical figure of Moses.
<narr> Narrative : A relevant document includes any information concerning Moses and his deeds regarding the Israelites.
</top>

d'information est ensuite déterminé. Chaque juge propose une dizaine de besoins d'informations candidats utilisés pour interroger le corpus. Un certain nombre de ces besoins d'informations seront sélectionnés notamment sur la base d'une estimation du nombre de documents pertinents présents dans le corpus pour ce besoin. Ces besoins d'information sont exprimés par des *topics*. Un *topic* comprend un numéro, un champ *title*, un champ *description* qui fournit une description du sujet de ce besoin d'information, un champ *narrative* qui donne les caractéristiques des documents attendus en réponse à ce besoin d'information. Les *topics* choisis sont utilisés par les participants qui renvoient chacun une liste de résultats fournis par leur SRI, constituée du classement de 1000 documents au plus. Ces documents sont utilisés pour former un *pool*. Chaque juge humain décide ensuite de la pertinence des documents du *pool* pour les topics qu'il a proposés. La figure 3.1 donne un exemple de *topic* de *TREC9* lié au corpus *WT10G*.
La suite de la construction consiste à fusionner la liste des n premiers documents retournés par chaque SRI pour former un *pool*. Les jugements de pertinence sont établis suivant un ensemble de lignes directrices fournis aux juges humains. Les documents non jugés (car non inclus dans le *pool*) sont considérés comme étant non pertinents pour tous les *topics*. Cette méthode a été décrite par *Sparck-Jones et van Rijsbergen* en 1975 [83] :

> *Idéalement, les jugements de pertinence doivent être exhaustifs. Si ce n'est pas le cas, des techniques doivent être utilisées pour effectuer des recherches indépendantes qui utilisent divers outils et informations afin d'obtenir un pool plus large de jugements de pertinence que ce qu'on obtiendrait en faisant simplement une évaluation utilisateur sur des résultats standards de moteurs de recherche.*[31]

[31] Ideally, these relevance judgements should be exhaustive. But if not some attempt should be made to carry out independant searches using any available information and device, to obtain a pooled output for more broadly based relevance judgements than may be obtained only with simple user evaluation of standard search output. In this case some estimate of the recall sample should be attempted.

Pour ou contre le *pooling* ?

Comme dit précédemment, les expérimentations ont toujours fait partie du domaine de la RI, et avec elles les critiques portant sur cette tradition expérimentale (*Cuadra et Katter* [37], *Harter* [64], *Taube* [152]). Le centre des critiques porte sur la subjectivité de la pertinence ; les jugements de pertinence varient d'un juge à un autre, et un même juge peut avoir des jugements de pertinence différents concernant la même information dépendant de son humeur, des pensées parasites ou de l'évolution de sa connaissance du sujet entre deux instants [138]. La première limite n'est donc pas celle de la technique du *pooling* en elle-même mais plutôt des collections de test : le jugement humain est au centre de la constitution de ces collections. Malgré les différents champs fournis dans les *topics* pour décrire plus précisément les documents attendus en réponse, il existe des différences entre les jugements de pertinence fournis par les juges humains parce que la pertinence est une notion subjective complexe à définir qui dépend de plusieurs variables. Ainsi, très tôt en RI, des travaux se sont penchés sur le taux de corrélation obtenu entre les classements des systèmes de recherche établis en utilisant différentes combinaisons de jugements de pertinence de différents juges. *Lesk et Salton* [96] ainsi que *Burgin* [25] n'ont pas trouvé de différences entre les performances relatives de diverses techniques d'indexation lorsqu'on fait une évaluation en utilisant différents ensembles de jugements de pertinence. *Cleverdon* [32] a également trouvé peu de différence dans l'ordre de 19 méthodes d'indexation en les évaluant par quatre ensembles (indépendants) de jugements de pertinence mais la corrélation entre les listes des méthodes était très élevée. De même la différence absolue en terme de performance entre ces méthodes était relativement faible. Les résultats de ces études ne sont pas généralisables aux collections utilisées de nos jours à cause de la taille réduite des collections utilisées à l'époque. Ainsi les travaux plus récents de *Voorhees* [155] ont étudié l'effet du changement de jugements de pertinence sur les évaluations réalisées avec les collections TREC. Les performances relatives de SRI sont stables malgré les changements effectués dans les jugements de pertinence. Le taux de corrélation élevé obtenu entre les classements des systèmes de recherche établis en utilisant différentes combinaisons de jugements de pertinence de différents juges montre que l'impact des désaccords des juges humains sur la classification finale des systèmes reste faible [32].

Une des limites majeures de la technique du *pooling* est liée à la croissance en volume des collections de test. Pour des collections vastes, l'incomplétude des jugements est de plus en plus élevée. La collection *Terabyte* qui comporte 25 millions de documents et occupe environ 426 giga-octets en est une illustration. Les tests ont été conduits pour déterminer la *complétude* des jugements

[32] Cette étude a tout de même détecté des conditions pour lesquelles la comparaison de systèmes doit être faite avec plus de précaution, notamment lorsque des topics ayant peu de documents pertinents sont pris en compte.

sur cette collection. Les résultats montrent que cette collection devrait être utilisée avec réserve, et l'utilisation de plusieurs métriques d'évaluation est préconisée ; il faudrait également tenir compte du nombre de documents non jugés retournés [31].
Les métriques classiques de recherche d'information ne sont pas robustes face à l'incomplétude des jugements de pertinence comme le montre les travaux de *Buckley et Voorhees* [24]. Une solution proposée consiste à adopter de nouvelles métriques plus fortement corrélées avec les métriques classiques quand il y a *suffisamment* de jugements de pertinence, mais plus robustes quand l'incomplétude des jugements augmente. Ces métriques sont exposées dans la section 3.4.1 de ce chapitre.
Dans la technique du *pooling*, les documents non jugés sont considérés dans la phase d'évaluation de systèmes de recherche comme étant des documents non pertinents. Or *Zobel* [165] montre qu'au mieux 50% à 70% de documents pertinents sont détectés en utilisant le *pooling*. Ces considérations risquent de pénaliser les *nouveaux* systèmes de recherche *i.e.* les systèmes n'ayant pas participé à la constitution du *pool* ou des systèmes utilisant des stratégies différentes de celles utilisées par les systèmes ayant participé au *pool*, et qui sont donc susceptibles de retourner des documents intéressants mais n'ayant été détectés par aucun système du *pool*. Les travaux de *Zobel* [165] se sont intéressés à la fiabilité des évaluations qu'il est possible de faire en utilisant les collections de test construites par *pooling*. Ces travaux concluent que malgré le biais introduit par l'incomplétude des jugements par le *pooling*, les jugements de pertinence fournissent une base crédible pour l'évaluation de nouveaux SRI (SRI n'ayant pas participé à la campagne *TREC*). Les travaux de *Kuriyama et al* [90] concluent également à une efficacité de la technique du *pooling* utilisée pour les collections de la campagne *NTCIR*.

Améliorations du *pooling*

Des améliorations de la technique du *pooling* ont été proposées mais des limites demeurent. Ainsi une variante des stratégies standards de *pooling* destinée à accroître le nombre de documents pertinents découverts a été proposée par *Zobel* [165]. Dans cette stratégie, le nombre de documents à juger par *topic* est incrémenté petit à petit (par pas de 10 par exemple) et des techniques de régression sont utilisées pour déterminer les *topics* susceptibles d'avoir encore des documents pertinents plus loin dans la liste des résultats. Les jugements de pertinence que ne sont produits que pour ces *topics* là. La méthode *Interactive Search and Judjing*, qui utilise un SRI interactif pour sélectionner les documents à juger et la méthode *Move-To-front Pooling* basée sur un nombre variable de documents pour chaque système participant au *pooling* suivant la performance de recherche de ce système est expliquée par *Cormack et al.* [35]. Une méthode de

pseudo-jugements de pertinence dans laquelle les juges humains sont remplacés par une technique aléatoire de sélection de documents pertinents est proposée par *Soboroff et al.* [144] et l'ordre des systèmes avec cette méthode est *positivement* corrélée avec leur ordre pour *TREC*. Améliorée, une telle méthodologie peut être utilisée pour des évaluations de moteurs de recherche du Web puisque le dynamisme du Web ne permet pas d'utiliser des techniques comme le *pooling*. *Sanderson et Joho* [130] proposent des techniques de construction de collections sans avoir recours au *pooling* qui permettent de former des collections aussi bonnes que celles de *TREC*. Des méthodes pour permettre la comparaison de SRI sans utilisation du *pooling* émergent également depuis peu (*Soboroff et al.* [144], *Wu* [163]).
Dans le cadre de campagnes d'évaluation comme *TREC*, des pistes de recherche spéciales sont dédiées à permettre l'évaluation de systèmes sur des collections volumineuses : la collection *Very Large Collection* [66] et la collection *TeraByte track* [4] de 426 Giga-octets ont ainsi vu le jour. Le tableau 2.4 du chapitre 2 donne un aperçu des diverses pistes de *TREC* depuis sa création et les tailles de quelques collections volumineuses utilisées.

3.3 Pertinence en RI

La pertinence est un concept qui n'est pas propre à la RI mais est commun à plusieurs domaines comme le montre *Greisdorf* [60] dand son aperçu de l'inter-disciplinarité de la pertinence. Elle est au centre de l'évaluation en RI. Elle détermine le degré auquel une réponse provenant du système de RI apporte satisfaction au besoin d'information de l'utilisateur. En effet, d'après *Belkin at al.* [17, 18], l'utilisateur d'un système de RI est dans une situation problématique qui nécessite des informations pour être résolue ; le niveau de pertinence donné à une unité d'information détermine le degré auquel elle participe à la résolution de ce problème.

3.3.1 Définitions - synonymes

Dès 1967, les travaux de *Cuadra et Katter* [37] ouvrent la boîte de Pandore de la pertinence. Plusieurs définitions ont été proposées depuis : *Goffman* [58] parle d'une mesure de l'information transportée par un document relativement à une requête et *Taube* [152] d'un prédicat psychologique qui décrit l'acceptation ou le rejet de la relation entre le sens ou le contenu d'un document et le sens ou le contenu d'une question.
Pour *Rees* [121], c'est un critère utilisé pour quantifier le phénomène qui a lieu quand des individus jugent la relation, l'utilité, l'importance, le degré de correspondance, la proximité, la convenance, la valeur des documents ou de leurs représentations par rapport à une demande

d'information, un besoin, une question ou une description d'une recherche. Cette définition englobe toute la complexité de la pertinence, complexité liée à ce qui est jugé précisément et à la base sur laquelle cette pertinence peut être établie. Ces points centraux sont présents dans les classifications des types de pertinence. Certaines des expressions qui ont été associées au concept de pertinence comme étant ses synonymes sont également présentées dans cette dernière définition. *Mizzaro* [109] présente plusieurs synonymes de la pertinence parmi lesquels *l'utilité*(*usefulness*) ; des expressions comme la *satisfaction de l'utilisateur*, le terme *topicality* sont souvent également associées à la pertinence. Quelques unes de ces expressions sont présentées comme suit :

utilité : (*usefulness*), elle couvre deux cas de figure : tout d'abord le cas où l'utilisateur a un besoin d'information donné, interroge le système de RI et parmi la liste des résultats, il obtient un document qui ne répond pas au besoin d'information exprimée mais qui répond à un autre besoin d'information de l'utilisateur : le document n'est pas pertinent mais il est utile. Le second cas est celui où un document retourné est pertinent pour le besoin d'information exprimé, mais ce document n'est pas utile (l'utilisateur connaît déjà l'information contenue dans le document). L'utilité est finalement une notion indépendante de la pertinence : un document peut être utile mais pas pertinent (pour le besoin d'information de départ), ou pertinent et pas utile, ni l'un ni l'autre ou les deux. Les travaux de *Kuhlthau* [89], *Sandore* [131], *Smithson* [143] cités dans *Spink* [147] abondent dans ce sens. Lors de l'évaluation du système de RI, il peut donc être intéressant de se pencher sur l'utilité des documents en parallèle à la pertinence.

La « topicalité » est une notion associée également à la pertinence. Dans le cadre des évaluations de SRI en recherche, l'on se base sur l'interprétation *thématique (topical)* de la pertinence en négligeant les autres aspects (*Cooper* [34]). Du point de vue des laboratoires, la *topicality* est aisée à définir et elle est utilisable/manipulable pour les systèmes car elle peut être reliée aux unités d'indexation (mots, texte). Un document peut être pertinent à différents degrés. La question du nombre de degrés à utiliser pour caractériser la pertinence n'a pas encore trouvé de réponse évidente en RI.

Finalement, la pertinence est une notion qui se comprend intuitivement (*Cosjin et Ingwersen* [45]) mais qui n'a pas vraiment de définition consensuelle, bien que de nombreux travaux s'y soient attelés. De *Cooper* [34] à *Schamber* [139], des formalisations et des classifications ont été proposées. *Cooper* [34] propose une définition basée sur la logique mathématique, *Wilson* [161] a amélioré cette définition en parlant de pertinence situationnelle (située) introduisant les objectifs de l'utilisateur. *Schamber et al.* [139] l'ont définie comme *un échange dynamique d'information qui dépend de la qualité de la relation entre l'information et le besoin d'information de l'utilisateur.*

Harter [63] cité par *Spink* [147] rejoint cette vision du dynamisme des jugements de pertinence en stipulant que les utilisateurs préfèrent des (informations) documents qui vont causer un changement de leur état cognitif, l'information étant finalement construite dans un processus constant. L'aspect dynamique de la RI par interaction est de plus en plus pris en compte pour l'évaluation (e.g. *Spink* [147], *Froehlich* [51]).

Saracevic [132] a proposé un cadre de classification des diverses notions de pertinence. *Froehlich* [51] propose un agenda sur la pertinence pour le 21ème siècle comprenant l'incapacité à définir la pertinence, la nature dynamique du comportement de RI[33] qui doit être étudié par des méthodologies appropriées et la nécessité de modèles cognitifs pour la conception des systèmes de RI et de leur évaluation. Les travaux de *Schamber* [138] parlent de trois points clés concernant la pertinence que sont le comportement (*behavior*), la façon de mesurer (*measurement*) et la terminologie de la pertinence.

Tous ces travaux montrent que la pertinence est en fait un phénomène social qui induit des processus cognitifs complexes ; un jugement de pertinence serait affecté par une quarantaine de variables d'après *Rees and Schulz* [122] alors que *Cuadra et Katter*[38] trouvent expérimentalement 38 variables parmi lesquelles le style, la spécificité, le niveau de difficulté des documents et que *Schamber* [138] en liste 80. La façon dont chaque critère affecte le jugement de pertinence de l'utilisateur n'est pas encore bien comprise (*Spink et al.* [147]). Finalement, il n'existe pas une notion unique de pertinence mais plusieurs pertinences affectées par divers critères.

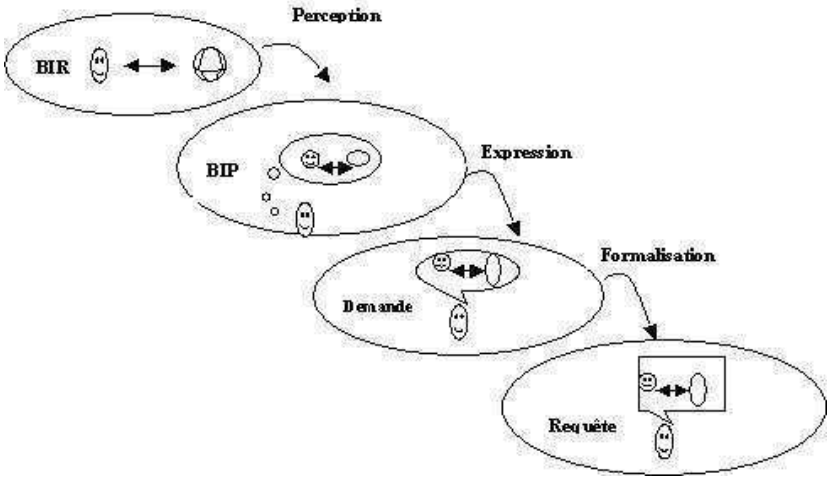
3.3.2 Pertinence : notion cognitive complexe et multidimensionnelle

La pertinence est un pont entre le besoin d'information et le résultat fourni par le système (ou plutôt la perception faite de ce résultat). Les étapes permettant de passer du besoin d'information réel de l'utilisateur à la requête finale soumise au système affectent donc la pertinence. Il en est de même pour les étapes pour passer du document à sa représentation par le système de RI. Un besoin d'information émane d'une tâche à laquelle il participe et se situe dans un contexte particulier ; tout ceci affectant aussi le jugement de pertinence de l'utilisateur. Pour *Mizzaro*, tous ces aspects donnent lieu à différents types de pertinence qu'on peut disposer dans un espace à 4 dimensions

- tout d'abord, les ressources d'information qui peuvent prendre différentes valeurs : document (entité physique retournée à l'utilisateur), *surrogate* (une représentation d'un document pouvant être son titre, une liste de mots-clés, etc.), information (l'entité non physique que reçoit ou crée l'utilisateur à la lecture des résultats de recherche).

[33] information seeking behavior

FIG. 3.2 – Etapes du besoin d'information de l'utilisateur [109]



- ensuite, la représentation du besoin d'information de l'utilisateur. Du besoin d'information réel (BIR) à la requête finale (exemple des mots séparés par des opérateurs booléens), l'utilisateur passe par différentes étapes :

  la perception qui donne lieu au besoin d'information perçu (BIP). La perception de son besoin d'information par l'utilisateur a été analysée par d'autres auteurs (*Mackay* [99] parle d'*incompleteness of the picture of the world*, *Belkin* [17, 18] parle d'*Anomalous state of knowledge* et *Ingwersen* [79] parle d'*Incomplete State of knowledge*).

  l'expression qui donne lieu à une demande de l'utilisateur. Cette étape souffre quant à elle du problème du vocabulaire (*Furnas et al.* [55]) (non correspondance entre les termes des documents et les termes de la requête, ambiguité des termes et synonymie).

  la formalisation dans un langage de requête (booléen par exemple). Elle peut être plus ou moins aisée selon le langage de requête.

  Une dernière étape éventuelle est celle de la représentation de la requête au sein du système de RI (lemmatisation, enlèvement de mots-vides, etc.). Le schéma 3.2 illustre ces différentes étapes.

  Les travaux de *Taylor* [153] ont proposé un cadre similaire pour formaliser le besoin d'information des utilisateurs[34].

- le temps : la dimension temporelle permet de prendre en compte l'évolution des connaissances de l'utilisateur au cours du temps ; en effet, pour un même besoin d'information à deux instants donnés, l'utilisateur n'aura pas la même satisfaction face à une réponse, en fonction de ce qu'il sait (ou ne sait pas) aux deux instants. Il est à noter que la connais-

[34] *Visceral need, conscious need, formalized need et compromised need.*

sance de l'utilisateur évolue aussi en fonction des résultats que lui renvoie le système (par exemple les premiers documents retournés). Pour *Wang et White* [160], le temps est une dimension cruciale. *Spink et al.* [147] ont également utilisé une dimension temporelle dans leur étude sur les régions de la pertinence, mettant ainsi en exergue son aspect dynamique (et l'aspect dynamique de la RI par interaction).

- La dernière dimension de *Mizzaro* [109] est nommée *composants* : une combinaison du *topic* (sujet d'intérêt de l'utilisateur), de la tâche (activité à exécuter par l'utilisateur avec les documents retournés) et du contexte.

Ainsi les différentes pertinences s'expriment par une fonction dépendant des quatre dimensions. Par exemple :
Pertinence(information, besoin réel, t, {*topic*, tâche, contexte}) est la pertinence de l'information réçue/créée par l'utilisateur à l'instant t pour le *topic*, la tâche pour laquelle il recherche de l'information et dans un certain contexte : c'est la pertinence qui intéresse l'utilisateur lambda. *Mizzaro* a positionné les types de pertinence proposés par d'autres auteurs dans son espace à quatre dimensions. Un des intérêts de cette formalisation est une meilleure prise en compte des divers aspects de la pertinence lors de la conception de systèmes de RI, en s'intéressant à ses différentes dimensions (voir les travaux de *Mizzaro* pour des exemples concernant la modélisation de la tâche ainsi que le présentation des résultats à l'utilisateur suivant chacune de ces dimensions). Les travaux de *Brooks* [22] parlent également de multidimensionnalité de la pertinence en créant une distance horizontale sémantique dans un espace multidimensionnel pour identifier « l'aura de la pertinence ».
Les jugements de pertinence qui sont définis par l'attribution d'une valeur de pertinence (par un juge) à un instant donné sont regroupés par *Mizzaro* [109] dans un espace suivant les 5 dimensions suivantes :

- type de pertinence jugée,
- catégorie des juges,
- les sources d'information que les juges peuvent utiliser,
- type de représentation du besoin d'information des utilisateurs auxquels les juges ont accès,
- le moment où le jugement de pertinence est exprimé

Les manifestations de la pertinence et ses attributs ont été étudiés par *Saracevic* [134] (cité dans (*Cosijn et Ingwersen*). Les différentes manifestations (ou niveaux d'après *Spink et al.* [147]) de la pertinence comprennent la pertinence système ou algorithmique, la pertinence thématique, les pertinences cognitive, situationnelle, motivationelle et affective. Le tableau 3.1 issu des travaux de *Cosijn et Ingwersen* [45] montre que les attributs de *Saracevic* (relation, intention, contexte,

inférence et interaction) peuvent s'appliquer à chaque type de pertinence[35].

3.3.3 Pertinence : binaire ou multivaluée

La division binaire de la pertinence est celle qui prévaut dans les travaux notamment pour la plupart des campagnes d'évaluation et de comparaison de SRI comme *TREC*. Un document est soit jugé pertinent soit jugé non pertinent par rapport à un *topic*, les documents non jugés sont considérés non pertinents lors du processus d'évaluation. Cette pertinence binaire à donné lieu aux métriques classiques de RI que nous présentons en section 3.4 mais elle n'est qu'une option parmi d'autres. D'après les constats effectués précédemment sur la complexité de la pertinence, il peut paraître réducteur de considérer que face à un besoin d'information, un document y corresponde ou pas. Il semblerait plus judicieux d'attribuer un degré auquel le document répond au besoin d'information. Cette vision est à la source de certains travaux de RI qui considèrent la pertinence comme une notion multivaluée.

La pertinence à plusieurs degrés n'est pas une notion récente. Dans les années 1960, *Hillman* [76] s'appuie sur les travaux de *Bar-Hillel* [12] (cité dans [60]) et affirme que les degrés de pertinence doivent être considérés en définissant une notion plus faible de parenté en termes de documents, questions et limites d'index[36]. *Fairthmore* dit son tracas face à la possibilité d'ignorer « la région du milieu exclue »[46]. Dans sa discussion sur la pertinence situationnelle, *Wilson* [161] admet également le souhait de reconnaître les degrés de la pertinence mais cet auteur n'étudie pas comment les mesurer. De même, pour *Wallis et Thom* [159] qui parlent de la nécessité de retourner tous les documents pertinents ou partiellement pertinents, mais stipulent aussi que les degrés de pertinence ne sont pas faciles à exprimer/formaliser.

Pour *Kekäläinen et Järvelin* [87], dans les environnements larges et modernes de RI, il est désirable d'avoir des SRI qui retournent des documents en fonction de leur niveau de pertinence. Par exemple les documents très pertinents sont retournés d'abord, ensuite les documents pertinents et enfin les documents faiblement pertinents. Les travaux de ces auteurs apportent des métriques d'évaluation basées sur une pertinence multivaluée, et dont la section 3.5 fera entre autres l'objet.

La région du milieu de la pertinence a été étudiée par *Saracevic et al* [137] qui donnent une

[35]Les mêmes auteurs ont proposé un modèle modifié de ce tableau dans lequel la pertinence affective est un attribut (*intentionality*) et pas un type de pertinence ; de plus ce nouveau modèle inclut un type de pertinence socio-cognitif.

[36]in defining a weaker notion of relatedness in terms of documents, queries and index terms

TAB. 3.1 – Attributs et manifestations de la pertinence [45]. U=utilisateur, I=information (les auteurs parlent de *information objects*)

Attributs	Manifestations				
	algorithmique	*thématique*	cognitive	situationnelle	motivationnelle
Relation	requête ⇒ I	*topic*⇒I	Etat de connaissance ⇒ I	situation, tâche perçue⇒ I	intentions, buts de l'U ⇒ I
Intention	dépendant du système intention derrière l'algorithme	expectation de l'U intention derrière la requête	Très personnel et subjectif lié au besoin d'information	très personnel subjectif ou émotionnel	
Contexte	Tuning de la performance de SRI(ex :TREC)	dépendant du contexte	dépendant du contexte	dépendant du contexte	dépendant du contexte
Inférence	fonctions de poids et de ranking	interprétation	processus individuel et subjectif	habilité de l'U à se servir de l'I en y donnant un sens	habilité de l'U à se servir de l'I en y donnant un sens
Interaction	retour de P automatique ou modification de la requête	les jugements de P sont dépendants du temps	très dépéndant du temps	inclut les interactions avec le domaine organisationnel et social	très personnel très personnel

définition de document partiellement pertinent[37].

Dans des travaux antérieurs, cette notion de niveau de pertinence est présente : *Pao* [117] et *Su* [149] ont utilisé trois niveaux de pertinence. *Tang et al.* [151] montre qu'une échelle à sept points est optimale en terme de confiance des utilisateurs dans leurs jugements de pertinence. *Howard* [77] a utilisé une échelle à 13 points et *Gluck* [57] une première échelle à cinq points et une deuxième à deux points. La collection du Web Track de *TREC* 2001 a également adopté une échelle à trois niveaux pour la pertinence des documents [157, 11]. Pour *Spink et Greisdorf* [147] l'existence d'une région du milieu de la pertinence est avouée ; cette région joue un rôle important pour les utilisateurs novices (notamment) en début de leur recherche et les amène peu à peu à un changement dans leur besoin d'information durant l'interaction avec le système de RI. *Spink et Greisdorf* ont décliné cette région en « pertinent », « partiellement pertinent »et « partiellement non pertinent »dans les jugements de pertinence fournis par des utilisateurs réels [147]. Ainsi, la pertinence est finalement vue comme une *relation* entre l'utilisateur et les résultats, elle est aussi un *effet* des résultats sur le problème de l'utilisateur et sur son processus de recherche. Ces auteurs ont proposé un modèle tri-dimensionnel (avec les trois dimensions que sont le type de pertinence, la region de pertinence et le temps), suggérant comme *Mizzaro* [109] la prise en compte de chacune de ses dimensions pour améliorer la conception de systèmes de RI.

Le nombre de degrés de pertinence varie donc selon les études, il varie aussi selon le niveau de détail qu'on souhaite atteindre, en fonction du contexte d'évaluation. La pertinence binaire correspond finalement simplement au cas où l'on souhaite donner le moins de détail sur le degré de pertinence d'un document. Le juste milieu à trouver entre une pertinence à deux valeurs et une pertinence à trop de valeurs dépend donc du contexte.

La pertinence à plusieurs degrés a toutefois eu quelques détracteurs comme *Cooper* [34] qui fournit une définition de la pertinence n'admettant aucun degré ou *Tiamyu et Ajiferuke* [154] qui argumentent pour un concept de « pertinence totale ». *O'Connor* [115] admet (sans l'étudier davantage) une distinction à faire entre des documents retournés pour satisfaire une requête clairement formulée et ceux offerts en lisant entre les lignes de la requête[38]. Dans bon nombre d'études parlant de pertinence à plusieurs degrés, les différents degrés de pertinence sont regroupés en une seule catégorie (pertinent) et concrètement dans les expérimentations, la pertinence est binaire.

La variété des travaux portant sur la pertinence montre la difficulté que les différents domaines des SI ont à définir clairement la manière dont les jugements de pertinence doivent être mesu-

[37] *Partially relevant : any document considered only somewhat, or in some part, related to the question or to any part of the question*

[38] volunteered as a result of reading between the lines of a request

rés. Il est vrai que la pertinence est finalement quelque chose qui émerge des interactions entre l'utilisateur (et son besoin d'information) et le système de RI. Le problème de la quantification de la pertinence est un problème ouvert en RI. Pour certains auteurs, elle ne serait même pas quantifiable/mesurable (*Foskett* [47]). Pour permettre des calculs sur des métriques, certains travaux ont (de façon empirique et arbitraire) attribué différentes valeurs numériques aux degrés de pertinence [88]. Nous présentons ces travaux en section 3.5. Dans nos travaux, nous avons formalisé des contraintes évidentes sur les degrés de pertinence et nous proposons des métriques d'évaluation basées sur la pertinence multivaluée (voir chapitre 5).

3.4 Evaluation par pertinence binaire

Les métriques d'évaluation classiques en RI s'appuient sur une notion de pertinence binaire. Dans bon nombre des travaux ayant pris en compte plusieurs degrés de pertinence, les degrés ont ensuite été regroupés en deux catégories « pertinent »et « non pertinent ». Nous présentons quelques métriques de RI couramment utilisées (dans les campagnes d'évaluation notamment). Des travaux de RI ont utilisé certaines de ces métriques pour déterminer le comportement des systèmes de RI quand ces derniers travaillent sur des ensembles de documents de plus en plus grands. Nos travaux présentés au chapitre 5 vont également dans ce sens. Face à une requête donnée, la pertinence binaire permet de subdiviser l'ensemble des documents d'une collection en quatre sous-ensembles tel que le montre la figure 3.3 ; les métriques à pertinence binaire peuvent s'exprimer en s'appuyant sur ces sous-ensembles.

3.4.1 Métriques classiques en RI

La précision et le rappel sont les métriques les plus souvent utilisées en RI. Elles ont été proposées par *Kent et al.* : le rappel est la proportion de documents retournés parmi ceux qui sont pertinents alors que la précision est la proportion de documents pertinents parmi ceux qui sont retournés. Ainsi, au travers de ces deux métriques, l'évaluation tient compte des documents pertinents détectés par la système de RI et des documents non pertinents refoulés par le système de RI (non retournés). Généralement, le rappel et la précision sont utilisés ensemble (sous forme d'une courbe rappel/précision construite généralement en calculant la moyenne de rappel/précision pour un système de RI pour un ensemble de requêtes). Les métriques moins connues existent, comme le *fallout* qui donne la proportion de documents non pertinents qui ne sont pas retournés ou la *généralité* qui est la proportion de documents pertinents au sein de la collection. La figure 3.4 donne des définitions de ces métriques. Quelques problèmes sont engendrés par

FIG. 3.3 – Répartition des documents d'une collection face à une requête (pertinence binaire).

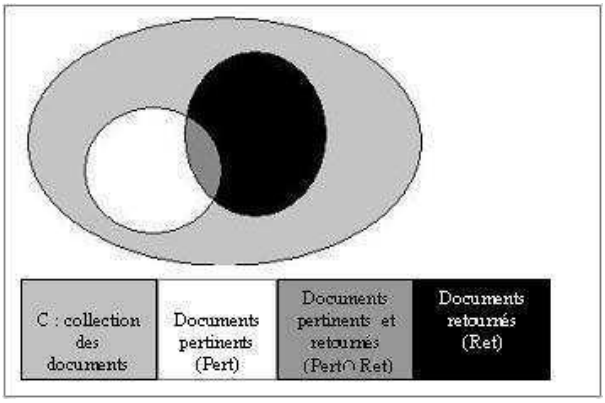


FIG. 3.4 – Métriques classiques de RI (pertinence binaire). Les différents ensembles sont définis en figure 3.3

Précision	Rappel	Fallout	Généralité
$\frac{\lvert Pert \cap Ret\rvert}{\lvert Ret\rvert}$	$\frac{\lvert Pert \cap Ret\rvert}{\lvert Pert\rvert}$	$\frac{\lvert C - Pert\rvert}{\lvert C - Ret\rvert}$	$\frac{\lvert Pert\rvert}{\lvert C\rvert}$

l'utilisation de ces métriques :

- pour le rappel, il est nécessaire de connaître l'ensemble de tous les documents de la collection qui sont pertinents par rapport à une requête ; dans les collections volumineuses, ceci s'avère impossible.
- l'utilisateur moyen est rarement intéressé par un taux de rappel élevé (autrement dit la connaissance de tous les documents pertinents), même s'il existe des applications pour lesquelles le taux de rappel s'avère important.
- une précision globale ne donne pas de renseignements sur la position des documents pertinents. La mesure de haute précision (précision sur les premiers documents retournés)[39] permet de pallier à ce problème. En effet, c'est une métrique fortement corrélée à la satisfaction de l'utilisateur moyen notamment pour la recherche sur le web ; de plus elle est facile à interpréter.
- la métrique de précision sur les premiers documents retournés ne permet toutefois pas de bien discriminer les systèmes de RI (elle s'occupe finalement uniquement de la présence ou non des documents pertinents parmi les premiers résultats). De plus, il est difficile d'interpréter une moyenne faite pour cette métrique pour différentes requêtes (elle est

[39]Par exemple la précision après 20 documents retournés notée généralement $P@20$ donne la proportion de documents pertinents parmi les 20 premiers documents retournés.

calculée après un nombre fixe de documents retournés, ce nombre correspond à différents niveaux de rappel selon les requêtes). D'après les travaux de *Buckley et Voorhees* [23], la précision après 10 documents retournés par exemple a une marge d'erreur plus grande que d'autres métriques (R-précision et *Mean Average Precision* notamment).

- Certains systèmes de RI ne fournissent pas un ordre total entre les documents retournés ; dans ce contexte, les méthodes de construction de courbes précision/rappel ne sont pas directement utilisables.

La précision et le rappel reflètent des aspects distincts et complémentaires des performances des systèmes de RI et sont donc utilisés ensemble, certains auteurs suggèrent l'utilisation d'une métrique unique pour l'évaluation. Des métriques combinant le rappel et la précision en une seule valeur ont ainsi vu le jour : moyenne harmonique, E-measure, etc.

Il existe également des métriques basées sur la précision et le rappel comme la R-précision (R étant le nombre de documents pertinents pour la requête considérée) ainsi que des histogrammes de précision requête par requête. La *R-precision* est une approche pour résoudre les problèmes de la précision après un nombre fixe de documents tel que présenté précédemment : chaque requête est évaluée au point où la précision et le rappel sont identiques. Sa marge d'erreur est plus faible que celle de $P@10$ mais plus grande que celle de la *MAP* [23].
La métrique *MAP* est la moyenne des précisions obtenues après chaque document pertinent retourné (on utilise zéro comme précision pour les documents pertinents non retournés). Elle correspond géométriquement à la région sous la courbe non interpolée de rappel/précision. Cette métrique s'appuie sur plus d'information sur la liste des résultats et est plus puissante et plus stable, mais pas aisée à interpréter.
Dans les différentes pistes des campagnes d'évaluation, il existe des métriques adaptées à la tâche/objectifs visés dans cette piste : exemple la *MRR (Mean Reciprocal Rank)*(moyenne sur l'inverse du rang du premier document pertinent) dans la tâche *Name Page Finding*. Les problèmes touchant la méthodologie de mise en place de collections de test affectent directement toutes ces métriques, particulièrement l'incomplétude des jugements de pertinence pour les collections volumineuses. De plus, ces métriques sont orientées système et ne prennent pas en compte l'utilisateur.
Elles ne sont pas robustes face à l'incomplétude des jugements de pertinence. Pour pallier à ce problème, les travaux de *Buckley et Voorhees* [24] proposent une métrique d'évaluation basée sur les positions relatives des documents pertinents et des documents *jugés* et non pertinents. L'idée est de mesurer les performances d'un système en se basant uniquement sur les documents jugés. Les métriques traitées précédemment ne font en effet aucune distinction entre des documents

jugés non pertinents suite au *pooling* et des documents non jugés. Cette nouvelle métrique se veut être une fonction du nombre de fois où un document jugé et non pertinent est retourné à un rang meilleur qu'un document jugé pertinent ; elle se base donc sur une notion de préférence d'où le nom *bpref* pour *binary preference.*

Soit un *topic* t ayant R documents jugés pertinents et d_{pert} un document pertinent pour t ; la *bpref* est calculée en fonction des documents d non pertinents pour t et retournés parmi les R premiers documents de la liste des résultats comme suit :

$$bpref = \frac{1}{R} \sum_{d_{pert}} 1 - \frac{d\,retourné\,avant\,d_{pert}}{R}$$

Une extension de la *bpref* est la *bpref-10* qui tient compte des *topics* ayant peu de documents pertinents.

$$bpref - 10 = \frac{1}{R} \sum_{d_{pert}} 1 - \frac{d\,retourné\,avant\,d_{pert}}{10 + R}$$

Cette métrique est corrélée aux métriques classiques quand des jugements de pertinence « suffisamment complets »[40] sont disponibles et elle est plus robuste à l'incomplétude des jugements de pertinence. Il est donc probable d'après ces travaux que de grandes collections de test construites en utilisant le *pooling* resteront des bases d'expérimentations fiables malgré l'incomplétude des jugements de pertinence.

Des métriques orientées utilisateur ont également été proposées [123] :

- ratio de couverture : fraction de documents connus de l'utilisateur comme étant pertinents et retournés
- ratio de nouveauté : fraction de documents pertinents retournés qui étaient inconnus de l'utilisateur.
- le rappel relatif : le ratio entre nombre de documents pertinents retournés et le nombre de documents pertinents espérés par l'utilisateur.
- l'effort du rappel qui donne le ratio entre le nombre de documents pertinents que l'utilisateur espérait trouver et le nombre de documents « lus »dans le but de trouver le nombre de documents pertinents espérés.
- la satisfaction et la frustration

Dans ces métriques, l'utilisateur est impliqué de façon partielle (les jugements de pertinence sont obtenus par des juges humains différents des utilisateurs, l'interaction entre l'utilisateur et le système pour construire son savoir n'est pas prise en compte).

[40] Comme le précisent *Kuriyama et al.* [90], le but de l'utilisation du pooling n'est pas d'obtenir tous les documents pertinents mais d'en obtenir suffisamment pour être capable de les utiliser et de faire une comparaison non biaisée des systèmes.

TAB. 3.2 – Métriques suivant le critère satisfaction de l'utilisateur noté U [148]

utilité	valeur(*worth*) des résultats vs temps passé valeur des résultats vs effort physique fourni valeur des résultats vs effort mental fourni valeur globale des résultats de la recherche
contribution de l'U	compréhension de la requête compréhension de la finalité (*purpose*) de la requête la complétude de l'U dans l'exploration de la question connaissance de l'U sur l'utilisation de la bd utilisation des termes appropriés pour le sujet de la requête
interview globale de l'U	succès global de l'interview
résultats de recherche	satisfaction de l'U sur la complétude des résultats confiance de l'U dans la complétude des résultats importance de la complétude des résultats satisfaction de l'U sur la précision de recherche importance de la précision des résultats
succès	jugement de l'U sur le succès global du SRI

Su [148] classifie vingt métriques en RI en utilisant les quatre critères importants que sont la pertinence, l'efficience, l'utilité et la satisfaction de l'utilisateur. Le tableau 3.2 donne quelques métriques pour le critère satisfaction de l'utilisateur. D'après les tests utilisateurs menés par *Su*, la métrique *valeur globale de la recherche* évalue au mieux la performance de la recherche et elle est fortement corrélée avec les métriques *satisfaction de l'utilisateur sur la complétude des résultats* et *satisfaction de l'utilisateur sur la précision de recherche*.

3.4.2 Evaluation du passage à l'échelle par des métriques utilisant la pertinence binaire

Les travaux de *Hawking et Robertson* [67] se sont intéressés à étudier le comportement des modèles de RI quand ces derniers font face à des espaces de recherche de plus en plus volumineux. Une étude théorique basée sur la théorie de détection de signal, et sur les distributions possibles de score de pertinence des documents pertinents et des documents non pertinents leur permet d'analyser l'impact de la taille sur les performances en RI, en utilisant comme métrique la haute précision. C'est la précision parmi les premiers documents retournés ; elle détermine la

FIG. 3.5 – Extrait des jugements de pertinence à trois degrés issus de la collection *WT10G*

topic		Document	Degré de pertinence
451	0	$WTX007-B50-87$	0
451	0	$WTX008-B17-23$	0
451	0	$WTX008-B26-172$	0
451	0	$WTX008-B37-10$	2
451	0	$WTX008-B38-114$	0
451	0	$WTX008-B39-477$	0
451	0	$WTX011-B03-195$	0
451	0	$WTX011-B16-71$	1
451	0	$WTX011-B28-183$	0
451	0	$WTX011-B28-222$	0

satisfaction de l'utilisateur dans des environnements comme le Web. Cette précision s'améliore avec l'augmentation du nombre de documents. Cette amélioration est justifiée par le nombre de documents pertinents plus élevé dans les échantillons plus grands, mais aussi dans la capacité du couple (requête, SRI) à classer les documents pertinents avant les documents non pertinents (puisque la précision diminue avec la taille de l'échantillon même dans le cas où elle n'est pas limitée par le nombre de documents pertinents présents dans l'échantillon).

3.5 Evaluation en RI par la pertinence multivaluée

3.5.1 Collections à pertinence multivaluée

WT10G de TREC9 : [11] Cette collection avait pour but d'être une collection de test pour des expérimentations concernant la RI sur le Web. Une échelle à trois degrés de pertinence a été adoptée (le tableau 3.5 donne un extrait des jugements de pertinence). Toutefois, les documents non jugés étaient aussi considérés non pertinents. Cette collection est utilisée pour nos expérimentations ; nous la présentons plus en détail en au chapitre 4.

INEX : La tâche générale d'un système de recherche dans des documents XML est définie comme celle de retourner, non pas des documents entiers, mais des composants de documents (éléments XML) qui sont le plus *spécifiques* et le plus *exhaustifs* par rapport au besoin d'information de l'utilisateur. Dans la campagne d'évaluation INEX, la pertinence

est une notion complexe représentée par ces deux aspects que sont : l'*exhaustivité* (qui reflète à quel degré un composant de document parle du sujet de requête) et la *spécificité* (qui reflète combien un composant discute de façon exclusive du sujet de la requête). Chacun de ces aspects influent sur la pertinence finale d'un élément XML : plus exhaustif et plus spécifique un élément est, plus cet élément est souhaité comme réponse par l'utilisateur. Chacun de ces deux aspects peut prendre plusieurs valeurs (comprises dans l'ensemble {0, 1, 2, 3}). Ces deux dimensions peuvent ensuite être combinées pour fournir une seule échelle de pertinence en utilisant une fonction de quantification comme suit :

$$f_{quant}(e, s) \ : ES \longrightarrow [0, 1],$$

A à chaque couple de valeur d'exhaustivité et de spécificité, cette fonction associe une valeur dans l'intervalle [0, 1] pour décrire l'ensemble des préférences de l'utilisateur, qui est le degré de pertinence de l'élément en question. Diverses fonctions de quantification sont utilisées selon que l'on souhaite privilégier l'exhaustivité ou la spécificité des éléments à retourner (*Kazai* [86]).

NTCIR Les campagnes NTCIR ont pour objectif de permettre des études sur la RI dans des documents en japonais. Trois degrés de pertinence ont été adoptés : pertinent, partiellement pertinent et non pertinent.

Collection à l'Université de Tampere : Dans le cadre d'un projet à l'Université de Tampere, les jugements de pertinence pour 38 *topics* issus des collections de *TREC-7* et *TREC-8* ont été refaits. Le but de ce projet était de créer une collection pouvant être utilisée pour étudier les documents très pertinents, la capacité des systèmes à les retourner et les différences de performances d'un système lorsqu'il est évalué en environnement à pertinence binaire par rapport à son évaluation au sein d'un environnement à pertinence multivaluée. Quatre degrés de pertinence ont été adoptés pour cette collection afin de mieux distinguer les documents marginalement pertinents des documents pertinents et des documents très pertinents. En effet, l'utilisation de trois niveaux de pertinence dans des collections comme la *WT10G* présenté précédemment ne permet pas cette séparation (*Sormunen* [146]). L'échelle à quatre points utilisée pour les jugements de pertinence est celle utilisée pour la construction d'une collection en finlandais au sein de l'Université de Tampere. Elle donne des indications sur ce qui est attendu d'un document de chaque niveau de pertinence (*Sormunen* [145]). Par exemple, le critère pour un document jugé *faiblement pertinent*[41] est qu'il comporte des informations à propos du besoin d'information de l'utili-

[41] *marginally relevant*

sateur mais qu'il comporte si peu d'information qu'il contribue difficilement à solutionner le problème de l'utilisateur[42]. Cette collection sera disponible courant 2007.

Terabyte : Cette piste de *TREC* voit le jour en 2004 et elle utilise la collection appelée *GOV2*. Les juges ont utilisé une échelle à trois degrés de pertinence (non pertinent, pertinent, hautement pertinent). Les juges humains déterminent eux mêmes les critères de distinction entre documents pertinents et documents très pertinents. Cependant, lors de la phase d'évaluation, les métriques à pertinence binaire sont utilisés, ce qui fait perdre le bénéfice que pourrait apporter ces degrés de pertinence (*Clarke et Scholer* [31]).

3.5.2 Métriques à pertinence multivaluée

Les principales métriques prenant en compte la pertinence multivaluée sont issues des travaux de *Kekäläinen et Järvelin*. Dans chacune de ces métriques, chaque degré de pertinence est représenté par une valeur numérique (arbitraire). Une des questions est le choix de ces valeurs et le sens dont elles peuvent être porteuses. Par exemple le niveau de pertinence « *hautement pertinent* »est représenté par la valeur numérique 3 et le niveau « *marginalement pertinent* »par la valeur 1 ; est-ce à dire qu'un document hautement pertinent est 3 fois plus pertinent qu'un document marginalement pertinent ? Cette question du choix des valeurs numériques à associer aux niveaux de pertinence et celle du nombre de degrés de pertinence restent ouvertes en RI. Parmi les métriques prenant en compte plusieurs degrés de pertinence, il y a celles qui étendent les métriques classiques comme la précision et le rappel et celles qui introduisent un nouveau concept : celui du gain d'information réalisé en fonction du degré de pertinence du document. Toutes ces métriques nécessitent de transformer l'aspect qualitatif d'un niveau de pertinence en une valeur quantitative. Les travaux ont généralement adopté une transformation empirique et *subjective*. Nous discutons de cette transformation au chapitre 4.

Attribution d'une valeur numérique à chaque niveau de pertinence

Dans les travaux qui gardent les degrés de pertinence jusqu'à la phase d'évaluation (comme dit précédemment, la plupart des travaux regroupent les différents degrés de pertinence et pour la phase d'évaluation la pertinence redevient binaire), il est incontournable d'attribuer une valeur numérique à chaque niveau de pertinence afin de réaliser les calculs des valeurs des métriques d'évaluation. Les travaux de *Kekäläinen* [88] basés sur quatre degrés de pertinence ont proposé quatre schémas de pondération différents, chaque schéma de pondération est censé refléter la

[42]On peut noter l'utilisation des termes comme *peu* ou *difficilement* qui sont des notions subjectives.

FIG. 3.6 – Quatres schémas d'attribution de valeurs numériques aux degrés de pertinence [88]

	très pertinent	pertinent	*faiblement* pertinent	non pertinent
schéma 1	1	1	1	0
schéma 2	3	2	1	0
schéma 3	10	5	1	0
schéma 4	100	10	1	0

valeur que l'utilisateur est supposé donner aux documents des différents degrés de pertinence (voir le tableau de la figure 3.6).

Extension des métriques classiques : *Generalized non binary precision and recall*

Kekäläinen et Jarvelin [87] proposent les métriques *generalized non-binary recall and precision* (précision et rappel non binaires) notées gP et gR qui sont des extensions des métriques classiques (précision et rappel) prenant en compte plusieurs niveaux de pertinence. Soit R l'ensemble des n documents retournés et appartenant à l'ensemble des documents ayant un niveau de pertinence pour une requête donnée, $R \subseteq D$, l'ensemble de tous les documents ; chaque document d a le niveau de pertinence $NivPertinence(d)$ qui est un nombre réel de l'intervalle [0..1], on pose

$$gP = \sum_{d \in R} NivPertinence(d)/n$$

et

$$gR = \sum_{d \in R} NivPertinence(d) / \sum_{d \in D} NivPertinence(d)$$

Comme les métriques de précision et de rappel classiques, ces métriques permettent des moyennes sur l'ensemble des requêtes, des moyennes de précision à des niveaux de rappel, des courbes de performances. Pour des environnements où les niveaux de pertinence ne sont pas compris dans l'intervalle [0..1], ils peuvent être ramenés à cet intervalle par une normalisation (en faisant le rapport avec le plus grand niveau de pertinence par exemple).

Cumulated Gain et Discounted Cumulated Gain

Les métriques *Cumulative Gain* et *Discounted Cumulative Gain* sont également des métriques qui prennent en compte plusieurs niveaux de pertinence et proposées par *Järvelin et Kekäläinen* [80]. Pour une collection donnée et un SRI, ces métriques calculent le gain cumulé d'information pertinente qu'on réalise au fur et à mesure qu'on parcourt la liste des résultats retournés par ce SRI sur cette collection. Dans le cas de la métrique *Discounted Cumulative*

Gain, l'information pertinente à un rang donné est pondérée par une fonction décroissante du rang, avant d'être cumulée. Pour chacune de ces métriques, un vecteur qui donne pour chaque rang l'information pertinente cumulée du premier rang jusqu'à ce rang est obtenu. Ces vecteurs peuvent être comparés à des vecteurs de gain d'information cumulée pour le cas d'un SRI *idéal* (SRI qui retourne les documents dans le meilleur ordre de niveau de pertinence, pour chaque requête). On peut soit comparer à un rang donné r le gain cumulé par rapport au gain cumulé idéal, soit visualiser les gains cumulés à chaque rang sous forme d'une courbe. A un rang donné r, le gain d'information cumulé est la somme cumulée des degrés de pertinence des documents de rang inférieur à r, soit :

$$\begin{cases} CG(1) = I(NivPertinence(d_1)) \\ CG(i) = CG(i-1) + I(NivPertinence(d_i)) \quad \text{si} \quad i \neq 1 \end{cases}$$

La métrique *Discounted Cumulative Gain (DCG)* calcule également des gains d'informations mais un coefficient de pondération (atténuation) qui est une fonction décroissante du rang est appliquée.

$$\begin{cases} DCG(1) = I(NivPertinence(d_1)) \\ DCG(i) = DCG(i-1) + NivPertinence(d_i) \times^b \log(i) \quad \text{si} \quad i \neq 1 \end{cases}$$

La visualisation des métriques que nous proposons s'appuie sur ce même principe et nous l'expliquons en détail plus loin. Les métriques proposées par *Sakai* [125] sont basées sur cette même notion de gain cumulé.

Extended CG et autres métriques dérivées de CG

Un des objectifs visés par par la campagne d'évaluation *INEX* est la définition de métriques d'évaluation pour la recherche d'information dans les documents XML. L'ensemble des métriques proposées utilise une valeur de pertinence unique qui combine l'exhaustivité et la spécificité, comme expliqué dans la section 3.5.1. Cette valeur est calculée par des fonctions différentes selon la métrique ; certaines convertissent la pertinence en une notion binaire (cas de la fonction dite d'agrégation *stricte* f_{strict} présentée ci-dessous, qui est la plus stable [86]) et d'autres obtiennent une pertinence à plusieurs degrés.

$$f_{strict}(s,e) = \begin{cases} 1 & \text{si} \quad e = e_{max} \, et \, s = s_{max} \\ 0 & \text{sinon} \end{cases}$$

s et e sont les valeurs de spécificité et d'exhaustivité pour un élément, et s_{max} et e_{max} sont les valeurs attribuées à la spécificité et l'exhaustivité maximales.

La fonction agrégation *généralisée* propose une échelle de valeurs de pertinence entre 0 et 1. *Exemple de la campagne de* 2005 (s est un réel entre 0 et 1 et e un entier égal à 0, 1 ou 2), l'agrégation généralisée était calculée par le produit de deux valeurs

$$f_{gen}(s, e) = s \times e$$

Les métriques *Cumulative Gain* et *Discounted Cumulative Gain* ont été *adaptées* à l'évaluation des SRI qui travaillent sur des documents structurés (XML). Dans le cadre de la campagne *INEX*, les métriques *XCG et nXCG ((normalized) eXtended Cumulative Gain)* sont utilisées [86]. Cette mesure du gain cumulé étendu donne le degré de pertinence, calculé par une des fonctions présentées ci-dessus, et accumulé au fur et à mesure de la progression dans la liste ordonnée des éléments retournés.

$$xCG(n) = \sum_{i=1}^{n} xG(i)$$

où $xG(i))$ est le degré de pertinence pour l'élément classé en position i par le système qu'on évalue. La comparaison entre ce gain cumulé et celui que le système aurait atteint s'il avait produit une liste triée *idéale* est faite pour obtenir le gain cumulé étendu normalisé.

3.6 Synthèse

Dans ce chapitre, notre intérêt à porté sur l'évaluation qui est une étape importante pour la mise en place des systèmes d'information. Ses concepts clés pour le domaine de la RI ont été définis et la notion de pertinence qui est au centre de l'évaluation en RI a été discutée. Tantôt concept binaire, tantôt concept multivalué, la pertinence induit des processus cognitifs complexes et peut être vue sous diverses dimensions. Les travaux ayant étudié l'influence du passage à l'échelle sur les performances des systèmes de RI, en utilisant des métriques de RI à pertinence binaire ont aussi été présentés. Le chapitre 4 porte sur nos travaux qui abondent dans ce sens.

Des métriques prenant en compte la pertinence multivaluée ont également été présentées. Sur ce point, nos conceptions rejoignent celles de *Kekäläinen et al.* [87] en ce qui concerne la prise en compte de multiples degrés de pertinence. Face à la masse importante et grandissante d'information, il existe des applications pour lesquelles les outils de recherche devront retourner en tête de liste de leurs résultats des documents du meilleur degré de pertinence. Toutefois, comme le montrent les travaux de *Spink*, il existe également des applications pour lesquelles il est utile de retourner des documents d'un degré moyen de pertinence (entre autres, ces travaux

stipulent que les documents très pertinents ne servent pas toujours à l'utilisateur moyen qui démarre sa recherche).

Finalement, il est donc judicieux de mettre en œuvre des techniques pour *classifier* les systèmes de RI en fonction de leur capacité à retourner des documents en fonction du niveau de pertinence de ceux-ci. Pour ce faire, il est indispensable de prendre en compte une notion de pertinence non binaire mais plutôt à plusieurs degrés. Dans le chapitre 5, nous concevons des métriques permettant d'évaluer la capacité des SRI à retourner une liste de documents triée par degré de pertinence décroissant, quand la taille des collections augmente.

Chapitre 4

Échantillonnage des collections pour étudier le passage à l'échelle

Sommaire

Dans ce chapitre, le premier axe de nos travaux concernant la construction de collections pour l'étude du passage à l'échelle est expliqué. Pour observer l'influence que la taille des collections a sur les modèles de recherche d'information, il nous a semblé opportun de disposer de collections de taille croissante sur lesquelles le comportement des modèles serait observé. Toutefois, il ne suffit pas d'avoir des collections de tailles différentes pour tirer des conclusions sur l'influence de la taille ; il est indispensable de contrôler le rôle du contenu même de ces collections sur les

performances des systèmes de recherche d'information. Nous proposons deux méthodologies pour mettre en place des ensembles de collections de taille croissante en réalisant ce contrôle. Chacune de ces méthodologies vise à réduire ou à contrôler le biais induit par le contenu des collections. Un cas pratique concernant l'évaluation en RI a permis d'étayer chacune de ces méthodologies. L'impact de la taille de collection sur les performances des SRI a ainsi été analysée en s'appuyant sur les métriques classiques d'évaluation en RI.

4.1 Introduction

Les chapitres précédents de cette thèse ont porté sur la problématique du passage à l'échelle, notamment dans les différentes phases d'un processus de RI. Des travaux ayant fourni des techniques pour observer l'influence de la taille de collection sur le comportement des modèles de RI ont été présentés. Ce chapitre présente notre approche concernant cet axe. Nous avons proposé deux méthodologies pour permettre d'étudier l'influence du passage à l'échelle sur les modèles de RI. Chacune d'elle consiste à partir d'une collection volumineuse et à construire un ensemble de sous-collections de taille croissante sur lesquelles le comportement des modèles de RI sera observé. Dans les deux techniques, le point de départ la détermination des *caractéristiques* de la collection volumineuse. Un contrôle est fait en cours de construction sur les caractéristiques de chaque sous-collection, celles ci doivent être *similaires* aux caractéristiques de la collection initiale. Cette contrainte affranchit du biais lié à l'impact de ce que contient chaque sous-collection et permet de focaliser uniquement sur la taille.

La première méthodologie présentée en section 4.2 est une démarche expérimentale reproductible (pour l'étude de l'influence du passage à l'échelle sur les modèles de RI) basée sur la construction d'une collection sur laquelle une caractéristique donnée C est la même quelle que soit la portion de collection sélectionnée. Cette nouvelle collection dite *uniforme* peut être découpée en sous-collections qui sont des *échantillons* de taille croissante de la collection entière et sur lesquelles des propriétés de modèles de RI sont étudiées. Cette démarche est appliquée de façon expérimentale sur la collection *WT10G* de *TREC9* avec comme caractéristique C la répartition des documents pertinents et comme propriétés les métriques d'évaluation de RI.

La deuxième méthodologie (section 4.3) utilise une technique d'échantillonnage connue en statistique. Nous utilisons en effet la méthode *Monte-Carlo* pour constituer des sous-collections de taille croissante ayant les mêmes caractéristiques (celles qui nous intéressent) que la collection initiale. Pour chaque taille, nous constituons un grand nombre de sous-collections de cette taille, ce qui de permet de calculer des intervalles de variation sur les valeurs numériques des métriques

donnant les performances des systèmes de RI et de fournir également une représentation graphique des performances des systèmes de RI au travers de boîtes à moustaches (*box and whiskers plots*). Grâce à ces intervalles de variation et aux boîtes à moustaches, nous pouvons tirer des conclusions ayant des bases statistiques. Nous avons également appliqué notre démarche de façon expérimentale sur la collection *WT10G* de *TREC9* avec comme caractéristique C la proportion des documents pertinents et comme propriétés les métriques d'évaluation de RI.
La dernière section (4.4) de ce chapitre donne une synthèse des travaux présentés et introduit le prochain chapitre.

4.2 Méthodologie d'échantillonnage par uniformisation

4.2.1 Nos hypothèses

Notre objectif est d'étudier l'influence de la croissance de collections sur les modèles de RI (notamment sur les performances de ces modèles). Une piste évidente consiste à observer le comportement de ces modèles sur des collections de taille croissante. Il ne suffit pas cependant d'avoir des ensembles d'informations de taille différente pour tirer des conclusions sur le comportement des modèles. Il est nécessaire d'avoir une certaine *logique* dans la croissance des collections ; ce qui *change* fondamentalement d'une collection à l'autre doit se résumer à la taille de collection, les autres *caractéristiques* de la collection restant les mêmes. Si ces conditions sont respectées, alors nous pouvons analyser l'influence de la taille de collections en observant les modèles de RI sur chaque collection. Une collection a évidemment de nombreuses caractéristiques qui peuvent la décrire, tout comme le comportement des modèles de RI peut être observé de différents points de vue. Il faut donc déterminer le point de vue suivant lequel on souhaite observer les modèles de RI. C'est ce que nous appelons les propriétés P_i des modèles qu'on veut étudier. En fonction du point de vue choisi (et donc des P_i), nous déterminons alors les caractéristiques de collections qui agissent sur ces propriétés. Ce sont ces caractéristiques qui vont être maintenues inchangées (identiques) d'une collection à l'autre (au cours du processus de croissance de la collection).
De plus, il ne suffit pas d'avoir une succession de collections de tailles différentes sans lien entre elles. Il s'agit de modéliser une *croissance* de collection (par exemple une collection C_1 qui grandit pour devenir une collection C_2, donc $C_1 \subseteq C_2$).

4.2.2 Méthodologie

Soit C une caractéristique de collection et P_i des propriétés sur lesquelles cette caractéristique a une influence. Le but de notre méthodologie est d'obtenir une collection sur laquelle la

caractéristique C est la même quelle que soit la portion sélectionnée de la collection. Si nous disposons d'une telle collection, il est possible de la découper en portions (sous-collections) de taille croissante, d'étudier des propriétés P_i sur chaque portion et d'analyser l'influence de la taille de la portion sur ces propriétés. Le découpage de la collection initiale est sans contrainte. Ceci signifie qu'on pourra découper la collection initiale de différentes manières, la seule contrainte étant que la caractéristique C choisie soit la même sur toute portion choisie. Notre méthodologie comprend 4 étapes :

1. Une première étape suppose que nous avons une collection initiale. Dans ce cas nous étudions la caractéristique C sur cette collection dans le but de déterminer si elle satisfait déjà la contrainte (*i.e.* si quelle que soit la portion de la collection, la caractéristique est la même). Si nous n'avons pas de collection initiale, nous débutons à l'étape 2 qui suit. Si une collection initiale qui satisfait les contraintes est disponible alors l'étape 2 est inutile et l'étape 3 est la suivante.
2. Une seconde étape (éventuelle) consiste à construire une collection sur laquelle la caractéristique C soit la même sur toute portion choisie.
3. La troisième étape est celle de l'échantillonnage de cette collection en portions de taille croissante (sous-collections).
4. Le dernière étape consiste à étudier les propriétés P_i sur chaque portion et analyser l'influence de la taille de la portion sur ces propriétés.

4.2.3 Cas d'étude

Cette méthodologie générale est appliquée au cas de l'évaluation en RI. Le chapitre 3 de cette thèse a montré la place centrale de la pertinence dans l'évaluation en RI. Ainsi, les métriques utilisées pour mesurer les performances des systèmes de RI sont basées sur les documents pertinents retournés (leur proportion, leur position, etc). Une étude de l'influence du passage à l'échelle sur l'évaluation de modèles de RI doit donc tenir compte des documents pertinents. Le cadre est celui de la pertinence binaire pour appliquer notre méthodologie à l'évaluation en RI. Ainsi, la caractéristique C est la répartition de documents pertinents et les propriétés à observer sont les performances des systèmes de RI, mesurées à l'aide des métriques d'évaluation de RI.

Hypothèses pour le cas d'étude

Quand les proportions de documents pertinents entre une collection et ses sous-collections ne sont pas les mêmes, il est délicat de mesurer l'impact qu'un accroissement de la taille aurait sur

les performances de SRI quand celles-ci sont évaluées en utilisant des métriques basées justement sur les documents pertinents. Si les documents pertinents sont répartis de façon uniforme sur l'ensemble de la collection, ceci permet de faire un découpage de la collection en sous-collections sans contrainte. Pour les hypothèses liées aux métriques classiques d'évaluation de la RI, l'analyse va porter, dans le cadre du passage à l'échelle, sur certaines propriétés de ces métriques que des travaux antérieurs ont mis en exergue. Par exemple, la métrique *MAP* (*Mean Average Precision*) est une métrique dite stable du fait qu'elle utilise plusieurs positions dans la liste de résultats retournés par un SRI [24].
Les quatre étapes de la méthodologie générale se déroulent comme suit :

Etape 1 : Etude de la caractéristique C

La première étape de notre méthodologie consiste à étudier la propriété C sur la collection initiale. Dans notre cas d'étude, une collection initiale est disponible. Il s'agit d'étudier la distribution de documents pertinents sur cette collection. Si les documents pertinents ne sont pas répartis de façon à ce qu'en sélectionnant n'importe quelle portion de collection, on en obtienne la même distribution, l'étape suivante est l'étape 2, sinon l'étape suivante est l'étape 3 directement.

Etape 2 : Construction d'une collection uniforme sur C

Cette étape a pour but d'obtenir une collection qui peut être découpée de différentes façons, en respectant nos hypothèses sur la caractéristique C. Pour notre cas d'étude, la distribution des documents pertinents doit être la même quelle que soit la portion de collection choisie. Ainsi, le nombre de documents pertinents par *topic* et pour tous les *topics* est proportionnel à la taille de la portion. Pour obtenir une telle distribution, la distance $E(t)$ souhaitée entre deux documents pertinents d'un *topic* t est calculée (cette distance donne en fait le nombre de documents non pertinents entre deux documents pertinents pour t). Soit $Pert(t)$ l'ensemble des documents pertinents pour le *topic* t, T l'ensemble de tous les *topics* et D l'ensemble de tous les documents de la collection. On a :

$$E(t) = \frac{\mid D \mid - \mid \bigcup_{k \in T} Pert(k) \mid}{\mid Pert(t) \mid}$$

Ainsi au sein de la nouvelle collection, les documents pertinents pour le *topic* t seront séparés de $E(t)$ documents dits non pertinents (*i.e.* documents qui ne sont jugés pertinents pour aucun *topic*), et éventuellement de documents jugés pertinents pour d'autres *topics* différents de t. Pour des documents jugés pertinents pour plusieurs *topics*, ils sont insérés une seule fois dans la

nouvelle collection, à la position définie par le premier des *topics* concernés traités[43]. Ainsi, la distance réelle entre deux documents pertinents de t notée $E_r(t)$ est telle que :
$E_r(t) \leq E(t) + \mid (\bigcup_{k \in T} Pert(k)) - Pert(t) \mid$.

Ceci introduit donc un biais sur l'uniformité qu'on souhaite avoir sur la collection. Étant donné que le nombre total de documents dans la collection est très élevé par rapport au nombre de documents pertinents ($\mid \bigcup_{k \in T} Pert(k) \mid << \mid D \mid$), ce biais est peu significatif et n'influence pas l'uniformité de la distribution des documents pertinents établie sur la collection. Lors du découpage de la collection en portions (sous-collections), la taille de ces sous-collections reste suffisamment élevée pour que l'influence de ce biais soit négligée. De plus, ce biais est un compromis nécessaire pour obtenir une collection au sein de laquelle les documents pertinents sont à la fois répartis uniformément (ou à peu près) pour chaque *topic* et répartis uniformément (ou à peu près) si l'on considère l'ensemble des *topics*. L'algorithme général de notre démarche d'uniformisation est donné par la figure 4.1.
Nous étayons cette démarche par un exemple.
Supposons que la collection initiale soit composée de 30 documents $D = \{d_1, \cdots, d_{30}\}$ et de deux *topics* $T = \{t_1, t_2\}$, avec $Pert(t_1) = \{d_1, d_7, d_{18}\}$ et $Pert(t_2) = \{d_7, d_{21}\}$. Le document d_7 est pertinent pour t_1 et pour t_2.

$E(t_1) = \frac{|D| - |\bigcup_{k \in T} Pert(k)|}{|Pert(t_1)|} = \frac{(30-4)}{3} \approx 8$ et $E(t_2) = \frac{(30-4)}{2} \approx 13$
Dans la collection uniforme, les documents sont ordonnés comme suit :

$$\{d_2, d_3, d_4, d_5, d_6, d_8, d_9, d_{10}, \underbrace{d_1}, d_{11}, d_{12}, d_{13}, d_{14}, d_{15}, \overbrace{d_7}, d_{16}, d_{17}, d_{19},$$
$$d_{20}, d_{22}, d_{23}, d_{24}, d_{25}, \underbrace{d_{18}}, d_{26}, d_{27}, d_{28}, d_{29}, d_{30}, \overbrace{d_{21}}\}$$

De façon théorique, la distance entre deux documents pertinents de t_1 est de $E(t_1)$ documents non pertinents. Dans le cas extrême pour l'exemple précédent, cette distance est de
$E(t_1) + \mid (P(t_1 \bigcup P(t_2)) - P(t_2) \mid$ soit $E(t_1) + 1$.

On remarque que le document d_7 est inséré une seule fois dans la collection. La distance réelle entre deux documents pertinents du *topic* t_1 est différente de la distance attendue parce que ce document est pertinent pour plusieurs *topics*. Le biais que cela implique sur l'uniformité de la collection n'est pas significatif si celle-ci est volumineuse.

[43] Ceci empêche de répliquer plusieurs fois le même document. En effet, pour des modèles de RI utilisant des paramètres comme l'*idf* ou le *tf*, la replication de documents peut avoir une influence sur les performances.

FIG. 4.1 – Algorithme de création de la collection uniforme

PertiEcrit=FAUX ;
$\forall t \in T$ {
i Initialiser compteur(t) ;
i Calculer E(t) ;
i Constituer Pert(t) ;
}
$NP = \{d \in D / \forall t \in T, d \notin Pert(t)\}$, $P = \bigcup_{t \in T} Pert(t)$
Tantque($NP \neq \emptyset$)et($\exists d \in P / NonMarquer(d, P)$){
i $\forall t \in T$ {
i Si (compteur(t)==E(t)) {
i d=document de Pert(t) ;
i Si (NonMarquer(d,P) {Ecrire(d) ;Marquer(d,P) ;}
i }
i Re-Initialiser compteur(t) ;
i PertiEcrit=VRAI ;
i }
i Si (PertiEcrit==FAUX){
i d=document de NP ;
i Ecrire(d) ;
i $\forall t \in T$, compteur(t)=compteur(t)+1 ;
i }
}

Etape 3 : Echantillonnage de la collection

Nous obtenons avec la démarche expliquée ci-dessus une collection *uniforme* et réutilisable pour différents types d'expérimentations. Cette collection peut être découpée en portions de tailles différentes et de différentes façons, puisque le nombre de documents pertinents sur une portion est proportionnel à sa taille.

Pour l'exemple précédent, pour obtenir des sous-collections de taille croissante, il est possible de découper la collection en $N = 3$ portions de taille $\frac{|D|}{N} = 10$

- Une première portion est $D_1 = \{d_2, d_3, d_4, d_5, d_6, d_8, d_9, d_{10}, d_1, d_{11}\}$
- Une seconde portion est $D_2 = \{d_{12}, d_{13}, d_{14}, d_{15}, d_7, d_{16}, d_{17}, d_{19}, d_{20}, d_{22}\}$
- Une troisième portion est $D_3 = \{d_{23}, d_{24}, d_{25}, d_{18}, d_{26}, d_{27}, d_{28}, d_{29}, d_{30}, d_{21}\}$

Ainsi, l'on peut construire des ensembles de sous-collections comme suit :
$\{S_1 = D_1, S_2 = D_1 \bigcup D_2, S_3 = D_1 \bigcup D_2 \bigcup D_3\}$ ou $\{S_1 = D_2, S_2 = D_1 \bigcup D_3, S_3 = D_1 \bigcup D_2 \bigcup D_3\}$.
Deux ensembles constitués chacun de trois sous-collections de taille croissante sur lesquelles la distribution de documents pertinents est la même sont ainsi construits.

Étape 4 : Etude de l'influence de la taille de collection

Dans cette étape, les sous-collections de taille croissante sur lesquelles la caractéristique C est la même sont disponibles (dans notre cas d'étude la caractéristique est la distribution des documents pertinents). Les propriétés P_i peuvent donc être étudiées sur chaque sous-collection et le comportement de ces propriétés quand la taille de collection augmente peut être analysé.

4.2.4 Expérimentations

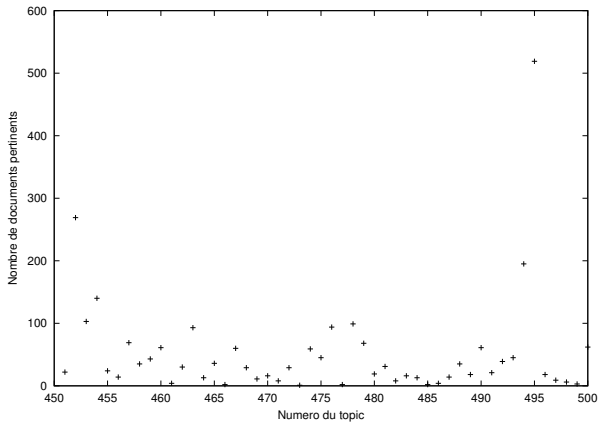
Données

Nous avons travaillé sur la collection de test de la conférence TREC dénommée WT10G [11]. Les besoins d'information pour nos tests correspondent aux *topics* 451-500 pour lesquels un ensemble de jugements de pertinence des documents est fourni. Cette collection de test contient 1 692 096 documents avec 2 617 jugements de pertinence pour les besoins d'informations 451-500 répartis comme le montre la figure 4.2.

La table 4.1 donne des statistiques sur les documents pertinents au sein de la collection WT10G.

$T = \{451, \cdots, 500\}$, D est WT10G et $\mid D \mid = 1\,692\,096$, $\mid \bigcup_{t \in T} Pert(t) \mid = 2\,588$. Ces 2 588 documents correspondent à 2 617 jugements de pertinence, avec 29 documents qui sont jugés

FIG. 4.2 – Nombre de documents pertinents par *topic* dans WT10G



TAB. 4.1 – Collection WT10G : statistiques sur le nombre de documents pertinents par *topic*

min	max	moyenne	médiane
1	519	52,34	29

pertinents pour plusieurs topics à la fois[44].

Les requêtes utilisées

A partir des *topics* de TREC9, nous avons construit un ensemble de requêtes pour l'interrogation des documents. Une requête est un ensemble de mots clés qui tiennent compte du titre du *topic* tel que fourni par TREC et du descriptif des documents pertinents attendus pour ce *topic*. Pour le *topic* 460 de TREC présenté comme exemple en figure 3.1 du chapitre 3, nous construisons la requête « *Israelites Moses bible biblical* ». Le tableau 4.2 donne des statistiques sur le nombre de mots-clés pour nos requêtes.

[44]Dans les expérimentations que nous avons menées, nous avons utilisé 2617 dans le calcul de l'écart $E(t)$ au lieu de 2588, ce qui correspond plutôt à $E(t) = \frac{|D| - \sum_{k \in T} |Pert(k)|}{|Pert(t)|}$; ceci a pour effet possible d'augmenter l'écart réel que nous aurons dans nos collections entre les documents pertinents d'un topic. Dans les cas extrêmes, on augmente cet écart de 29 documents. La plus petite taille de collection que nous avons utilisé est de 200 000 documents. Cette augmentation de l'écart est très faible face à la taille de nos collections, elle n'affectera pas l'uniformité que nous désirons, c'est pourquoi nous pouvons la négliger.

TAB. 4.2 – Statistiques sur nos requêtes : nombre de mots par requête

Min	Max	Moyenne
2	8	4,76

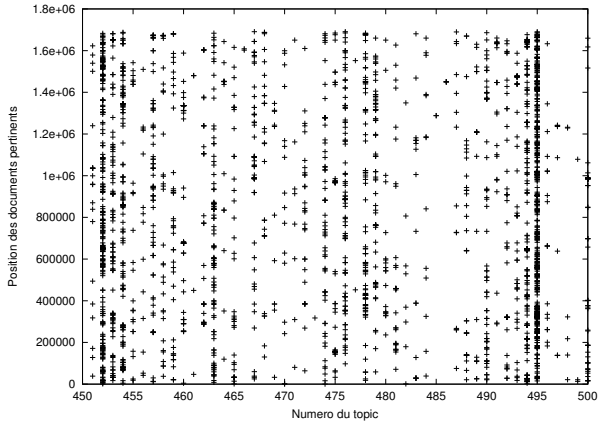
Modèles de RI utilisés

- L'outil LUCY [5] (désormais connu sous le nom de *Zettair* [6]) que nous avons utilisé implémente le modèle *BM25* de Okapi qui est une extension du modèle probabiliste.
- Nous avons également utilisé l'outil MG [7] qui est basé sur le modèle vectoriel (MG offre également le modèle booléen mais nous ne l'avons pas utilisé). Dans la version 1.2.1 que nous avons utilisée, MG ne gère, pour les termes, que les caractères sur un octet de l'ASCII. Pour les documents en Français, nous avons donc remplacé tous les caractères accentués par leur équivalent sans accent et, par un espace tous les autres caractères comme par exemple, le symbole ©.
- une implémentation du modèle Okapi que nous nommerons *Okapip* a été réalisée.

Les trois autres modèles (méthodes) sont basés sur la proximité entre termes de la requête. Brièvement présenté, voici le principe de chacune de ces méthodes ; les implémentations que nous avons utilisées sont décrites en détail dans les travaux de *Mercier* [106].

- La méthode *Cover Density Ranking* de *Clarke et al.*[30] permet de classer les documents pertinents du point de vue du système selon la densité de couverture des mots-clés de la requête dans les documents. Soit une requête ayant k termes. Tout document contenant au moins un mot de la requête est pris en compte. L'ensemble de ces documents est classé dans l'ordre décroissant du nombre de mots de la requête contenue dans le document. Pour chacun de ces documents, un calcul des *intervalles* du document contenant au moins deux termes de la requête est ensuite effectué. Ensuite, la méthode sélectionne d'abord les documents qui possèdent des intervalles contenant tous les k termes de la requête, puis les documents possédant des intervalles contenant les $k-1$ termes et ainsi de suite jusqu'à 1. Cette méthode sera nommée *modèle de Clarke*.
- La méthode de *Hawking et Thistlewaite.* [68] : Une requête est un ensemble de couples (u, a), composés d'une relation de proximité u et d'un coefficient d'importance a. Pour un document d, $I(u, d)$ est l'ensemble des intervalles de d qui satisfont la relation de proximité u. Chaque intervalle de $I(u, d)$ participe au score de pertinence de ce document. Cette méthode sera nommée *modèle de Hawking*.
- La méthode de *Rasolofo et Savoy* [120] attribue à chaque document, pour une requête donnée, un score calculé sur la base du modèle BM25 de Okapi et en fonction de la proximité des occurences des paires de termes de la requête dans le document. Cette méthode sera nommée *modèle de Rasolofo*.

FIG. 4.3 – Position des documents pertinents par *topic* (collection initiale)



Distribution des documents pertinents sur WT10G

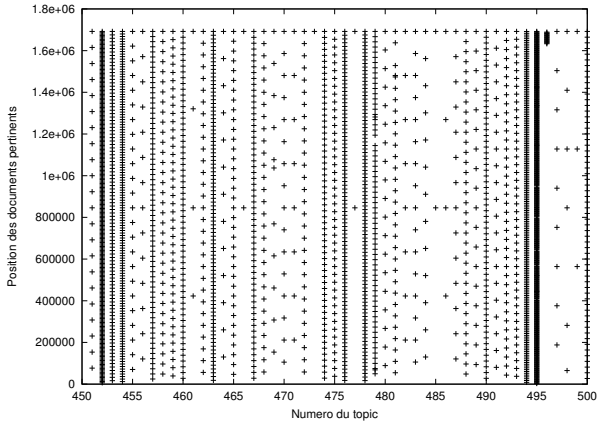
La répartition (position) des documents pertinents sur la collection WT10G par *topic* a été étudiée et nous obtenons le graphique de la figure 4.3. La répartition des documents pertinents est quelconque et elle varie selon les *topics*. Pour un *topic* donné, l'ensemble des documents jugés pertinents n'est pas réparti uniformément sur l'ensemble de la collection, le nombre de documents pertinents n'est pas une fonction linéaire de la taille de la collection de test. Etant donné que la collection initiale sera subdivisée en sous-collections de taille croissante, il est important de tenir compte de la façon dont sont répartis les documents sur chaque sous-collection pour que les métriques de RI gardent tout leur intérêt quelle que soit la taille de la collection utilisée. Une redistribution des documents pertinents au sein de la collection a donc été faite.

Uniformisation de la collection

La nouvelle répartition des documents pertinents obtenue sur la collection initiale est appelée une répartition uniforme. La figure 4.4 montre cette distribution uniforme. Dans cette nouvelle collection, le nombre de documents pertinents pour un *topic* donné est fonction linéaire de la taille de la collection et le nombre de documents pertinents (tous *topics* confondus) est également une fonction linéaire de la taille de la collection. Ainsi au sein de la nouvelle collection, les documents pertinents pour le *topic* t seront séparés de $E(t)$ documents dits non pertinents (*i.e.* documents qui ne sont jugés pertinents pour aucun *topic*), et éventuellement de documents jugés pertinents pour d'autres *topics* différents de t. Par exemple, le *topic* numéro 495 possède 487 documents jugés pertinents.

$E(495) = \frac{(1\,692\,096 - 2\,617)}{487} = 3\,469$ et $E_r(495) \leq E(495) + 2\,617$

FIG. 4.4 – Documents pertinents par *topic* (Collection uniforme)



TAB. 4.3 – WT10G Uniforme : statistiques sur les longueurs de documents en nombre de caractères par sous-collection (Documents pris dans l'ordre d'apparition après uniformisation)

Taille collection	Min	Max	Moy	Médiane
200000	2	2.326.791	3.875	1.623
400000	2	2.326.791	4.101	1.554
600000	2	2.326.791	3.902	1.529
800000	2	2.344.748	3.876	1.543
1000000	2	2.344.748	3.856	1.533
1200000	2	2.344.748	3.767	757
1400000	2	2.344.748	3.773	1.504
1600000	2	2.344.748	3.790	1.492
1692096	2	2.344.748	3.796	19.292

Découpage en sous-collections

Dans le cadre de nos expérimentations, des portions de taille croissante par pas de 200 000 documents ont été constituées, en prenant les documents dans l'ordre d'apparition dans la collection uniforme. 8 sous-collections décrites dans la table 4.3 ont été construites et 7 d'entre elles ont été utilisées. Le choix du pas peut varier selon les besoins des expérimentations. Il est possible de découper la collection entière en N sous-collections de taille k avec $k \times N$=taille de la collection entière) et ensuite mixer ces N sous-collections de différentes façons pour obtenir des collections de différentes tailles, puisque l'on sait que le nombre de documents pertinents sur une portion est proportionnel à la taille de celle-ci.

4.2.5 Résultats

Nota Bene

Dans les travaux présentés en [78], la collection WT10G avait été utilisée pour nos expérimentations. Certains résultats présentés dans cet article diffèrent de ceux présentés dans le cadre de cette thèse. En effet, dans les expérimentations présentés dans [78], nous avons utilisé comme documents pertinents uniquement ceux ayant le niveau de pertinence 1. Or les documents de la collection WT10G avaient été jugés suivant trois niveaux de pertinence (0, 1, 2), les documents ayant les niveaux de pertinence 1 et 2 constituant l'ensemble des documents pertinents. Dans les expérimentations présentées dans ce chapitre de thèse, l'ensemble de tous les documents jugés pertinents est pris en compte.

Il est à noter que le nombre de *topics* ayant effectivement été pris en compte pour chaque sous-collection varie (voir le nombre de *topics* donné dans le tableau 4.5). Ceci s'explique de deux façons : toutes les sous-collections ont été interrogées avec l'ensemble des 50 *topics* disponibles pour la WT10G. La première raison est liée à la distribution uniforme des documents. Les *topics* ayant peu de documents pertinents ou ayant plusieurs documents pertinents communs à d'autre *topics* n'ont pas forcément des documents pertinents sur toutes les sous-collections. La seconde raison est qu'il existe également des *topics* ayant effectivement des documents pertinents présents dans la collection mais certaines méthodes n'en retournent aucun. Dans le calcul, ces derniers *topics* ne sont pas pris en compte pour les méthodes concernées. Donc, pour une méthode et une sous-collection, les calculs sont faits sur l'ensemble des *topics* ayant au moins un document pertinent retourné par cette méthode sur cette sous-collection.

Parmi les modèles de RI utilisés, l'outil LUCY nous a permis d'indexer et d'interroger toutes les sous-collections et la collection WT10G uniforme entière. Pour des raisons techniques, les indexations et interrogations réalisées avec les autres méthodes s'arrêtent à la collection de taille 1 400 000 documents.

Courbes rappel/précision

Les courbes de rappel/précision pour les 6 modèles de RI sont données comme suit : modèle de *Clarke* (figure 4.5), modèle de *Hawking* (figure 4.6), modèle *Okapi de Lucy* (figure 4.7), modèle vectoriel de l'outil *MG* (figure 4.10) et modèle de *Rasolofo* (figure 4.9) et modèle *Okapip* (figure 4.8).

Pour tous les modèles de RI utilisés, on remarque que les courbes rappel/précision pour les grandes collections sont très similaires, sur l'ensemble des niveaux de rappel. Dans nos travaux

FIG. 4.5 – Rappel/précision pour les 7 sous-collections uniformes de WT10G - modèle de *Clarke*

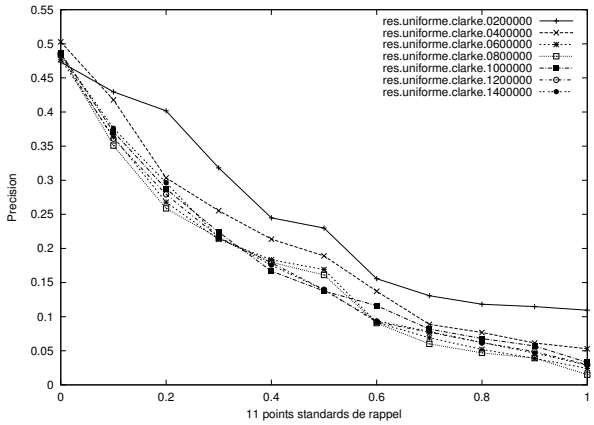


FIG. 4.6 – Rappel/précision pour les 7 sous-collections uniformes de WT10G - modèle de *Hawking*

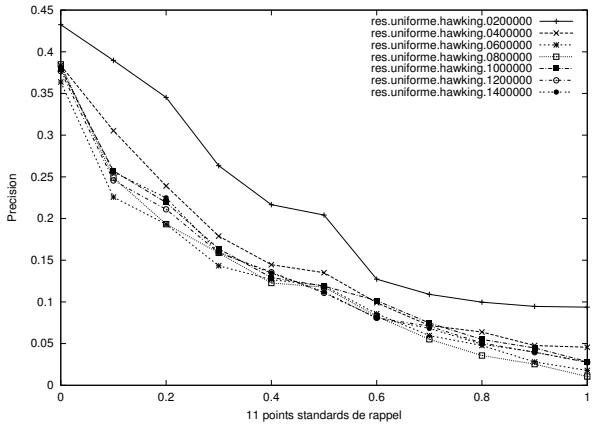


FIG. 4.7 – Rappel/précision pour les 7 sous-collections uniformes de WT10G - modèle Okapi(Lucy)

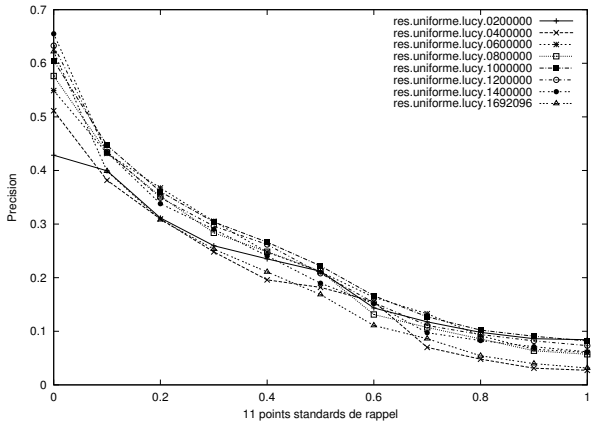


FIG. 4.8 – Rappel/précision pour les 7 sous-collections uniformes de WT10G - modèle Okapip

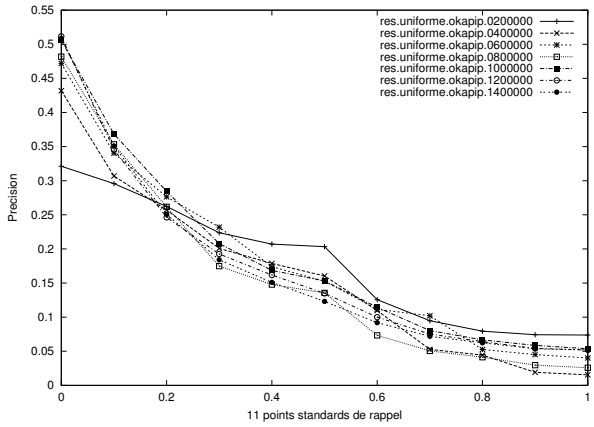


FIG. 4.9 – Rappel/précision pour les 7 sous-collections uniformes de WT10G - modèle de *Rasolofo*

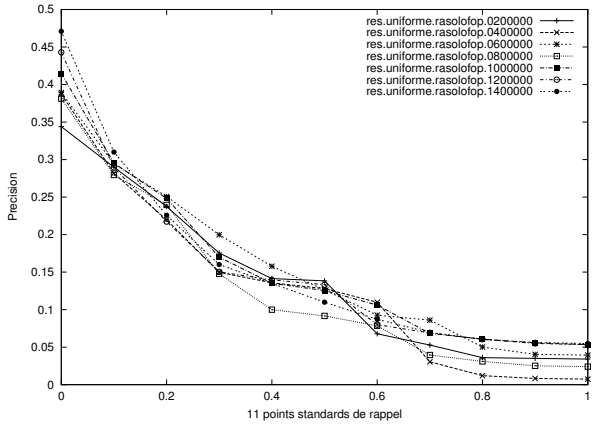
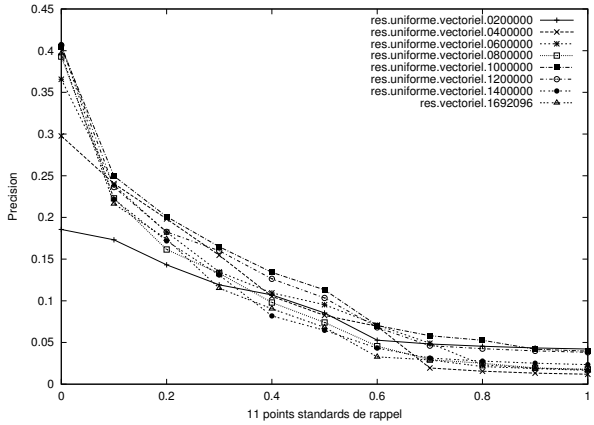


FIG. 4.10 – Rappel/précision pour les 7 sous-collections uniformes de WT10G - modèle vectoriel (Mg)



cités en [78], cette similitude entre les courbes des grandes collections était associée à une certaine stabilité des modèles. Les expérimentations présentées ici ne semblent pas aller dans ce sens. Il semble plus plausible que cette similitude soit liée à la façon dont croissent les sous-collections. Un pas unique a été adopté pour la croissance des collections et la quantité de documents nouveaux d'une sous-collection à la sous-collection suivante est donc la même à chaque fois. Toutefois, la proportion de nouveaux documents ainsi apportée n'est pas la même. Une collection de 400 000 documents à laquelle l'on rajoute 200 000 nouveaux documents se voit augmenter de 50% alors qu'une collection de 1 400 000 documents à laquelle l'on rajoute 200 000 nouveaux documents se voit augmenter de 1/7 seulement. La nouveauté apportée dans le second cas est faible et elle aura donc un impact moindre (autrement dit les résultats de la collection à 1 400 000 documents et ceux de la collection à ces 1 400 000 + 200 000 nouveaux documents seront proches). Dans le premier cas, la nouveauté apportée est grande et aura un impact plus grand. Il est donc normal d'avoir un rapprochement et une similitude entre les courbes de grandes collections par rapport à celles des petites collections.

Pour le modèle de *Clarke* et le modèle de *Hawking*, les courbes des petites collections sont au dessus des courbes des grandes collections ; la précision se détériore avec la croissance en taille des collections sur l'ensemble des niveaux de rappel.

Pour le modèle de *Rasolofo* et le modèle *Okapip*, les grandes collections dominent et sur les premiers niveaux de rappel, la précision s'améliore avec la taille de la collection. Sur les niveaux de rappel plus grands, il y a chevauchement entre les différentes courbes et aucune tendance particulière ne se dégage.

Quant au modèle vectoriel de *MG*, les grandes collections sont également dominantes sur les premiers niveaux de rappel et la précision s'améliore avec la taille de la collection.

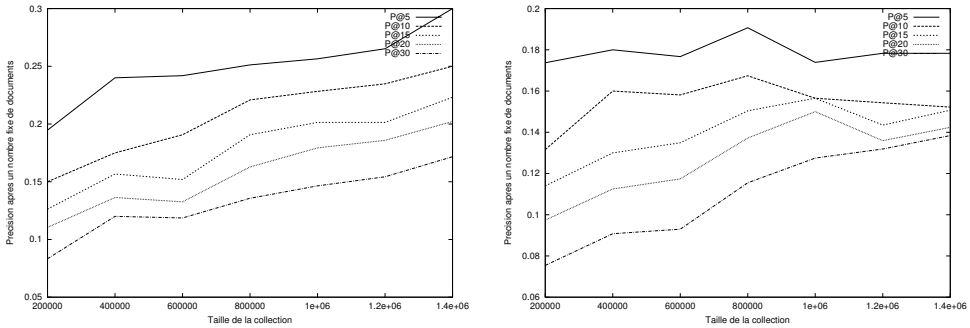
Pour le modèle *Okapi* de Lucy, on remarque également une dominance des grandes collections sur les petites, sur les premiers niveaux de rappel.

Analyses

Le modèle vectoriel (MG), le modèle de *Rasolofo*, le modèle *Okapip* et celui de *Lucy* tiennent compte de la taille de la collection dans leur technique de calcul de score des documents (dans le classement de *Hawking et Robertson*, ils sont en variance déterministe). Quant aux modèles de *Hawking* et de *Clarke*, ils ne tiennent pas compte de la (taille) de la collection, puisque le score d'un document dépend uniquement des termes de la requête et du document (dans le classement de *Hawking et Robertson* [67], ils sont classés dans les modèles à variance statistique).

Ainsi le rôle de la collection dans l'attribution de scores aux documents pourrait affecter le

FIG. 4.11 – Précision sur les 1ers documents retournés pour les 7 sous-collections - Modèle de *Clarke* et modèle de *Hawking*



passage à l'échelle du modèle (en ce qui concerne le rapport rappel/précision). La performance rappel/précision baisse avec l'accroissement des collections pour les modèles de RI pour lesquels le score attribué à un document dépend uniquement de la requête et du document. Pour les modèles pour lesquels le score dépend de la collection, c'est le contraire. Pour ces derniers modèles, l'estimation des paramètres du score qui dépendent de la collection pourrait s'améliorer sur les grandes collections.

Haute précision

La satisfaction de l'utilisateur dans des environnements comme le Web est directement liée à la précision sur les premiers documents retournés. De plus, cette précision est une métrique d'évaluation de SRI qui est facile à interpréter. Les travaux de *Hawking et Robertson* [67] ont utilisé un unique modèle de RI (Okapi BM25 au travers du système PADRE) pour mener des expérimentations sur la précision après un nombre fixe de documents. Pour les 5 modèles de RI utilisés, nos résultats rejoignent ceux que ces auteurs ont obtenu pour les $P@20$. Nous étendons ces résultats aux précisions $P@5, P@10, P@15, P@30$. Les figure 4.11 (modèle de *Clarke* et modèle de *Hawking*), figure 4.12 (modèle Okapi de Lucy et modèle Okapip) figure 4.13 (modèle de *Rasolofo* et modèle vectoriel de l'outil MG) montrent l'évolution de ces précisions quand la taille de la collection croît. Pour les 5 modèles utilisés, nous obtenons des courbes similaires.

Ces résultats montrent que la précision sur les premiers documents retournés augmente avec la taille de collection.
De plus, comme attendu de façon intuitive, nous avons $P@n >= P@m$ si $n < m$ [45]

Les travaux de *Hawking et Robertson* [67] stipulent que sous l'hypothèse de disposer de suf-

[45] Dans nos travaux [78], nous avions des courbes similaires mais l'analyse faite comportait quelques erreurs et avait conduit à conclure autrement.

FIG. 4.12 – Précision sur les 1ers documents retournés pour les 7 sous-collections - Modèle Okapi de Lucy et modèle Okapip

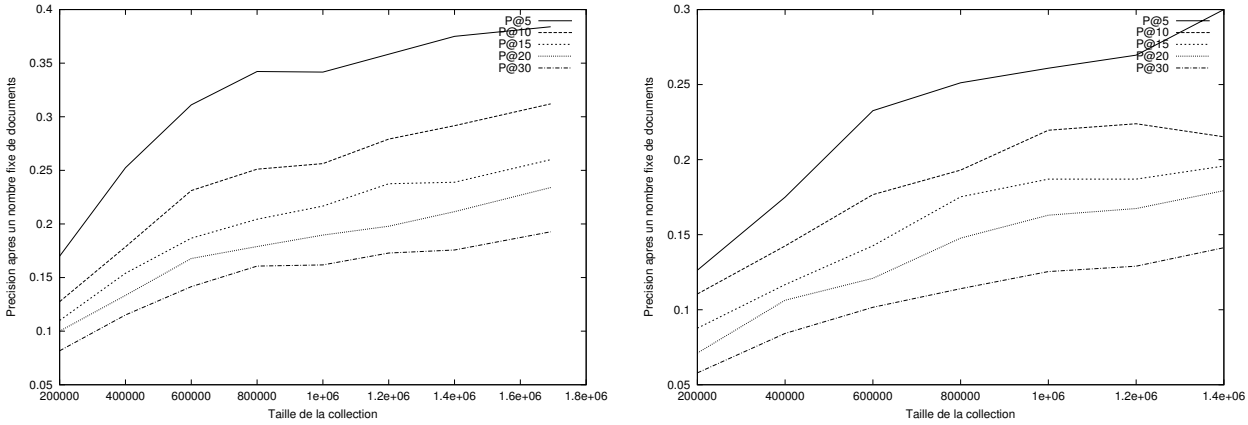
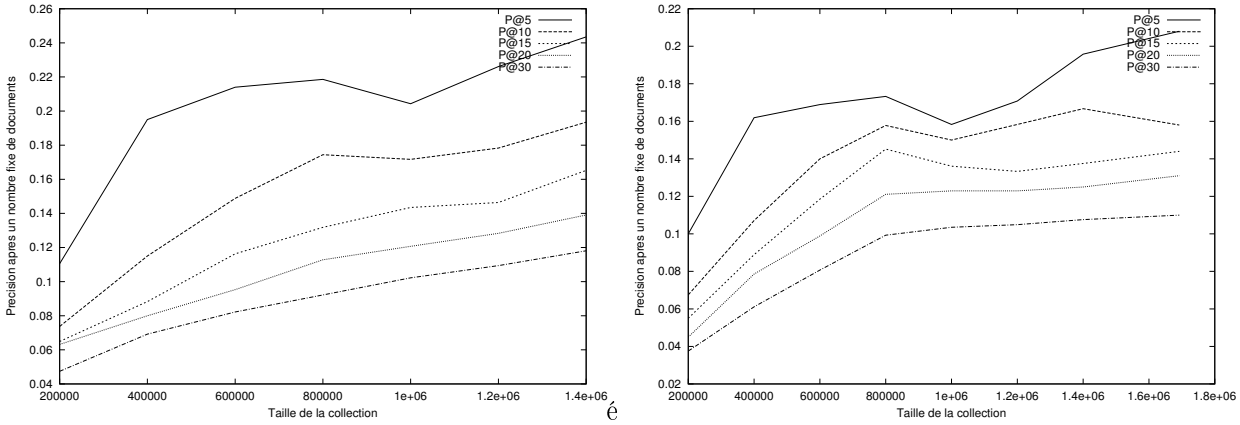


FIG. 4.13 – Précision sur les 1ers documents retournés pour les 7 sous-collections - Modèle de *Rasolofo* et modèle vectoriel de MG



é

TAB. 4.4 – Position moyenne du 1er document pertinent retourné. Nous donnons la moyenne en considérant les *topics* ayant un document pertinent retourné avant le rang 100

	Clarke	Hawking	Rasolofo	Okapip	Vectoriel	Lucy
200000	31	31	31	30	23	35
400000	36	36	32	33	29	37
600000	37	38	35	37	35	40
800000	38	38	34	36	35	40
1000000	39	37	37	41	36	42
1200000	39	37	37	37	36	43
1400000	40	39	38	41	38	43
1600000						44
1692096						46

fisamment de documents pertinents dans chaque sous collection, la précision après n documents retournés dans une collection de taille T devrait prédire la précision après $n*x$ documents retournés dans une collection de taille $T*x$. Ainsi, $P@5$ dans la collection de taille 200 000 documents pourrait prédire $P@20$ dans une collection de taille 800 000 documents ou prédire $P@30$ dans une collection de taille 1 200 000 documents. Nos résultats n'ont pas permis d'abonder dans le sens de cette observation, pour les différents modèles utilisés.

Positions des premiers documents pertinents

Le tableau 4.4 et le tableau 4.5 donnent des statistiques sur les positions des premiers documents pertinents retournés au fur et à mesure que la taille de collection augmente (moyenne et médiane). Une position de document faible indique un document retourné en tête de liste des réponses. La moyenne est donnée en considérant les *topics* ayant un document pertinent retourné avant le rang 100. La position moyenne du premier document pertinent est plutôt stable avec la croissance des collections, pour tous les modèles. La position médiane du premier document pertinent retourné reste stable et faible quelque soit la taille de la collection, ce qui signifie que pour au moins la moitié des *topics*, le premier document pertinent est en tête de la liste des résultats.

Le tableau 4.6 donne le taux de *topics* pour lesquels le premier document pertinent retourné est dans une position inférieure à 5 (colonne de gauche pour chaque méthode) et à 10 (colonne de droite). Globalement, le pourcentage de *topics* pour lesquels le premier document pertinent retourné est dans une position inférieure à 5 ou à 10 augmente avec la taille de la collection.

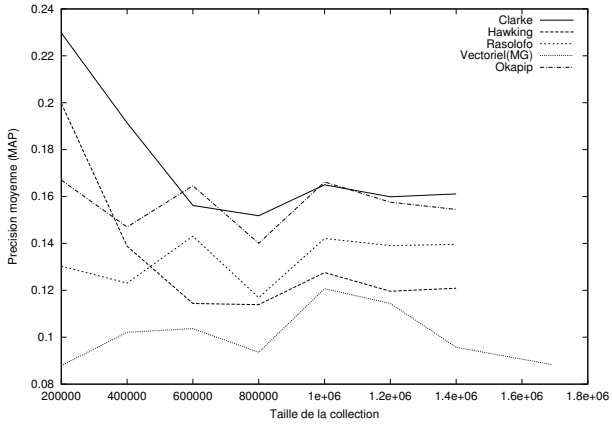
TAB. 4.5 – Position médiane du 1er document pertinent retourné. Le nombre de *topics* est également donné

	Clarke	Hawking	Rasolofo	Okapip	Vectoriel	Lucy
200000	2,5/38	3/38	16,5/38	3,5/40	64/40	2/40
400000	2/40	4/40	4,5/40	3,5/40	15/42	2/42
600000	3/43	5/43	4/43	3/43	7/43	2/45
800000	3/43	5/43	4/43	3/43	6/45	2/45
1000000	3/46	7/46	4/46	3/46	6/48	2/45
1200000	3/46	6,5/46	3/46	3/46	6,5/48	2/48
1400000	3/46	7,5/46	3/46	3/46	5,5/48	1,5/48
1600000						2/48
1692096						1/49

TAB. 4.6 – Pourcentage de *topics* pour lesquels le ler document pertinent retourné est en position <5 (gauche) et < à 10(droite)

	Clarke	-	Hawking	-	Rasolofo	-	Okapip	-	Vectoriel	-	Lucy	-
200000	65	76	60	68	36	42	36	47	27	34	52	67
400000	70	75	60	70	52	57	52	60	40	42	66	73
600000	67	72	55	65	55	62	62	69	44	51	68	77
800000	60	74	51	65	58	67	67	69	46	53	71	73
1000000	60	71	43	60	58	65	67	71	47	56	72	77
1200000	60	71	47	63	63	67	67	71	47	58	75	81
1400000	63	73	45	63	67	71	67	71	50	66	75	81

FIG. 4.14 – Précision moyenne (MAP) en fonction de la taille de la collection.



Courbes de MAP

Les courbes donnant l'évolution de la précision moyenne (MAP) en fonction de la taille de collection pour les différents modèles utilisés sont présentées sur la figure 4.14. Pour les modèles de *Hawking* et de *Clarke*, la précision moyenne baisse rapidement quand on part des petites collections, elle reste ensuite relativement stable sur les collections plus grandes. Pour les modèles vectoriel de *MG*, *okapip* et le modèle de *Rasolofo*, la précision moyenne oscille dans un petit intervalle de valeurs, sans que l'on puisse noter une croissance/diminution particulière avec l'augmentation de la taille de collection. La précision moyenne est une métrique très stable puisqu'elle se base sur plusieurs valeurs de performances sur toute la liste de résultats retournés par le SRI (en effet, elle utilise la précision après chaque document pertinent retourné). L'augmentation de la taille de collection ne l'améliore ni ne la détériore particulièrement.

Apports

Nous proposons une démarche expérimentale reproductible qui permet d'étudier l'influence du passage à l'échelle sur les modèles de recherche d'information. Cette démarche passe par l'uniformisation d'une collection volumineuse par rapport à une caractéristique C et son échantillonnage en sous-collections de taille croissante. Nous étudions ensuite, en utilisant ces sous-collections, l'impact de l'augmentation en volume du corpus sur des propriétés liées à la caractéristique C. Les résultats montrent que les modèles de RI pour lesquels l'attribution de score à un document ne dépend pas de la collection (mais uniquement du document et de la requête) semblent améliorer leurs performances rappel/ précision quand la taille de collection croît.
Les résultats pour la haute précision confirment (et étendent à plusieurs niveaux de coupure et à 5 modèles de RI) ceux de *Hawking et Robertson* [67], à savoir que la précision après un nombre

fixe de documents retournés augmente avec la taille de la collection.
Les résultats obtenus montrent également que la croissance en taille de la collection améliore véritablement la position du premier document pertinent retourné et que la variabilité de la MAP est faible face à cette croissance.
La croissance du volume d'information est continue et il devient incontournable de s'intéresser à la façon dont les modèles de recherche vont se comporter face à des espaces de recherche de plus en plus larges. Notre méthodologie vise à permettre de telles études. Elle peut être utilisée pour bâtir des collections « uniformes »par rapport à d'autres caractéristiques (répartition des termes de requêtes) pour étudier d'autres propriétés de RI.

Cette première technique met en place une collection uniforme qui peut être subdivisée en sous-collections. Sur les petites collections, il y a un risque qui est celui de n'avoir aucun document pertinent pour des *topics* ayant très peu de documents pertinents. De plus, dans l'étude de cas, l'utilisation d'une seule série de sous-collections fait que les résultats obtenus peuvent être fortement liés à cette série. Ceci n'est pas une limite liée à la méthodologie mais plutôt à l'utilisation que nous en avons fait. D'après la méthodologie, il est en effet possible d'avoir différentes séries de sous-collections tout en respectant la caractéristique de similarité C.
Il a semblé opportun de renforcer la méthodologie proposée en réduisant le risque évoqué ci-dessus. De plus, la création de multiples séries de sous-collections de taille croissante permettrait l'utilisation des techniques statistiques pour interpréter les résultats. C'est le but de la seconde méthodologie présentée dans la section suivante.

4.3 Méthodologie d'échantillonnage aléatoire

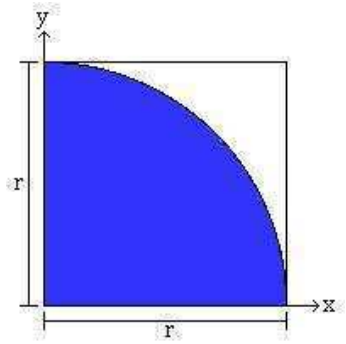
4.3.1 Quelques points de statistique

Méthode de Monte Carlo

On qualifie de méthode Monte-Carlo tout algorithme de calcul reposant sur des tirages aléatoires, dont le résultat change selon ces tirages et est d'autant plus précis que le nombre de tirages est élevé. Un exemple souvent utilisé est celui du calcul de la valeur approchée de Pi de la manière suivante : on prend des points au hasard (uniformément) dans le carré $[0,1] \times [0,1]$ et on regarde combien sont dans le quart de disque de centre O et de rayon r (voir la figure 4.15). Le rapport du nombre de points à l'intérieur sur le nombre total de points est proche du rapport de la surface du quart de cercle sur la surface du carré, soit $\pi/4$. On montre que :

$$\frac{Nombre\ de\ points\ dans\ l'aire\ grise}{Nombre\ de\ points\ dans\ le\ carre} \text{ s'approche de } \frac{1/4 \times \pi \times r^2}{r^2} = \frac{\pi}{4}$$

FIG. 4.15 – Quart de cercle utilisé pour déterminer la valeur de $\pi/4$ par la méthode de Monte Carlo.



lorsque le nombre de points augmente.
Une des mises en œuvre informatiques consiste à considérer que le cercle est de rayon $r = 1$. En générant de façon aléatoire deux nombres x et y compris dans l'intervalle [0, 1] et en utilisant le théorème de *Pythagore*, on obtient la distance de l'origine $(0, 0)$ (centre du cercle). Si cette distance est inférieure ou égale à 1 alors le point est dans l'aire grise. En répétant cette expérience un grand nombre de fois, on obtient une valeur de plus en plus proche de celle de $\pi/4$. La méthode de Monte Carlo est utilisée dans des domaines divers et variés comme l'économie ou la physique. Avec cette méthode, un système large peut être échantillonné en des configurations choisies de façon aléatoire, et ces configurations sont utilisées pour décrire statistiquement le système entier.

Box-plot **sur la distribution d'une variable**

Le terme *Box and Whiskers Plot* est une invention de *Tukey* en 1977 pour représenter schématiquement une distribution. De façon plus générique, le terme *Box Plot* est aussi utilisé. Ces deux termes recouvrent une grande variété de diagrammes en forme de boîtes. Les traductions francophones couramment utilisées sont celles de *boîte à moustaches* ou *boîte à pattes*. La représentation graphique que nous utilisons s'appuie sur la médiane et les quartiles concernant une série de valeurs ; nous rappelons les définitions de ces concepts par un exemple.
Soit la série de 9 valeurs données : 1, 3, 4, 5, 6, 7, 9, 10, 15

- La médiane est le point qui partage la série en deux groupes d'effectifs égaux ; elle est aussi nommée deuxième quartile. Pour notre série elle vaut 6,
- Le 1er quartile $Q1$ repartage le groupe du bas (valeurs inférieures à la médiane) en deux groupes d'effectifs égaux ; il vaut 4 pour notre exemple,
- Le 3eme quartile $Q3$ repartage le groupe du haut (valeurs supérieures à la médiane) en deux groupes d'effectifs égaux ; il vaut 9 pour notre exemple,

Selon que l'effectif de la série est pair ou impair, on procède différemment pour évaluer les quartiles. L'écart interquartile ($Q1 - Q3$) correspond à 50% des effectifs situés dans la partie centrale de la distribution.
Sur la boîte à moustaches d'une variable, on retrouvera l'échelle des valeurs de la variable, les valeurs des premier, deuxième et troisième quartiles. On retrouve aussi deux moustaches inférieure et supérieure qui délimitent les valeurs dites *adjacentes* qui sont déterminées à partie de l'écart interquartile. Les valeurs dites extrêmes, atypiques ou exceptionnelles situées au delà des valeurs *adjacentes* sont aussi représentées par des marqueurs particuliers.

Intervalle de confiance sur une moyenne

L'intervalle de confiance sur une moyenne μ fournit une fourchette d'estimation d'une moyenne à partir d'un échantillon. Le niveau de confiance et la taille de l'échantillon interviennent.
Soit à calculer l'intervalle de confiance de niveau $1 - \alpha$ d'une moyenne μ fondé sur la moyenne empirique $\hat{\mu}$ observée après une expérience portant sur n individus ; la variance observée s^2 est calculée et utilisée pour le calcul de l'intervalle par l'approximation normale usuelle (théorème de la limite centrée) :

$$IC_{(1-\alpha)\%} = [\hat{\mu} - z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}; \hat{\mu} + z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}]$$

C'est l'intervalle de confiance de niveau approché α sur la moyenne ou encore l'intervalle de confiance à $(1 - \alpha)\%$ sur la moyenne. L'approximation ci-dessus est jugée satisfaisante pour $n \geq 30$. $(1 - \alpha)\%$ est la probabilité à priori que les valeurs calculées à partir de l'échantillon incluent la vraie valeur de μ. $z_{1-\frac{\alpha}{2}}$ dépend de α, c'est le quantile à $1 - \frac{\alpha}{2}\%$ de la loi normale centrée réduite. Par exemple pour un intervalle de confiance avec le niveau de confiance 95%, soit $\alpha = 5\%$, $z_{1-\frac{\alpha}{2}}$ vaut $1,96$.

4.3.2 Nos hypothèses

Les hypothèses émises pour l'échantillonnage par la méthode Monte-Carlo sont proches de celles de la première méthodologie, en ce qui concerne les caractéristiques de la collection qui restent inchangées quand celle-ci croît en taille. L'utilisation de la méthode de Monte Carlo permet d'obtenir pour une taille donnée un grand nombre de sous-collections. Les performances de ces sous-collections sont recueillies et peuvent être utilisées pour établir des analyses statistiques par taille de sous-collection.

4.3.3 Méthodologie

Nous construisons des sous-collections de taille croissante. Pour une taille donnée, un grand nombre de sous-collections de cette taille est construit. La caractéristique C de chacune de ces sous-collections est identique à celle de la collection entière. Les étapes de cette méthodologie sont les suivantes :

1. Une collection initiale est disponible et la caractéristique C est étudiée sur cette collection.
2. Pour une taille donnée, un grand nombre de sous-collections est construit de cette taille en effectuant une sélection aléatoire des documents mais de façon à respecter la caractéristique C.
3. Pour une taille donnée, les propriétés P_i sont étudiées sur chaque sous-collection de cette taille et les propriétés P_i correspondant à cette taille sont déduites.
4. Les propriétés correspondant à chaque taille de sous-collections sont observées graphiquement (en utilisant par exemple des boîtes à moustaches) et l'on peut alors analyser l'influence de la taille sur ces dernières d'un point de vue statistique.

4.3.4 Cas d'utilisation et expérimentations

Une fois de plus, cette méthodologie a été appliquée au cas de l'évaluation en RI. Les propriétés P_i sont les métriques d'évaluation en RI (à pertinence binaire) et la caractéristique est la proportion de documents pertinents dans chaque sous-collection.

Collections, requêtes, modèles de RI

Les collections, les requêtes et les modèles de RI que nous avons manipulés dans cette partie expérimentale sont les mêmes que ceux utilisés pour les expérimentations de la première méthodologie.

Construction des sous-collections

Nous calculons la proportion à laquelle correspond la sous-collection à construire par

$$Prop = \frac{\mid D \mid}{\mid Taille(Sous - Collection) \mid}$$

Pour une taille donnée de sous-collections, pour chaque *topic* t, nous calculons le nombre de documents pertinents de ce *topic* à insérer dans toute sous-collection de cette taille, soit :

$$NbDocs(t) = Prop \times \mid Pert(t) \mid$$

TAB. 4.7 – Statistiques sur la longueur des documents par taille de sous-collection (30 sous-collections ont été construites pour chaque taille)

Taille sous-collection	min	max	moyenne	médiane
200000	2	2344748	3831	1497
400000	2	2344748	3826	1497
600000	2	2136913	3805	1500
800000	2	2344748	3802	1501
1000000	2	2344748	3806	1500
1200000	2	2344748	3798	1500
1400000	2	2344748	3792	1500

Il est à noter que tout document pertinent pour plusieurs *topics* est inséré une seule fois dans la sous-collection mais pour le compte de tous les *topics* concernés. Il est donc possible d'avoir un nombre total de documents pertinents inférieur à $\sum_{t \in T} NbDocs(t)$. Les documents pertinents pour un *topic* t sont choisis de façon aléatoire au sein de $Pert(t)$. Le nombre total de documents pertinents est complété par des documents non pertinents (qui ne sont pertinents pour aucun des *topics*), pour atteindre la taille souhaitée de sous-collection ; les documents non pertinents sont également sélectionnés de façon aléatoire.

Nous avons utilisé sept tailles de sous-collections pour nos expérimentations : 200 000, 400 000, 600 000, 800 000, 1 000 000, 1 200 000, 1 400 000 documents. La section suivante donne les statistiques sur les sous-collections construites pour chaque taille ainsi que les résultats obtenus par interrogation avec les modèles de RI présentés précédemment.

Dans les représentations des boîtes à moustaches, nous avons présenté les trois quartiles, les valeurs adjacentes et les valeurs atypiques.

Les intervalles de confiance présentés sont calculés avec un niveau de confiance de 95%.

4.3.5 Résultats

Quelques statistiques sur les sous-collections construites

Les tableaux 4.7 et 4.8 donnent des statistiques sur les sous-collections construites.

Description courbes rappel/précision

Pour chaque taille de collection, les courbes rappel/précision pour l'ensemble des échantillons de collection de cette taille sont tout d'abord présentées (cas de la méthode de *Clarke*). Les

TAB. 4.8 – Nombre théorique de documents pertinents attendus par groupe de sous-collections.

Taille sous-collection	Documents pertinents
200 000	309
400 000	618
600 000	927
800 000	1 237
1 000 000	1 546
1 200 000	1 855
1 400 000	2 165

courbes pour l'ensemble des échantillons d'une taille ont le même allure pour les autres modèles.

Les courbes rappel/précision pour les grandes collections sont très homogènes (voir les figures

FIG. 4.16 – Rappel/précision pour les 30 sous-collections de WT10G à 200 000 documents et à 400 000 documents- modèle de *Clarke*

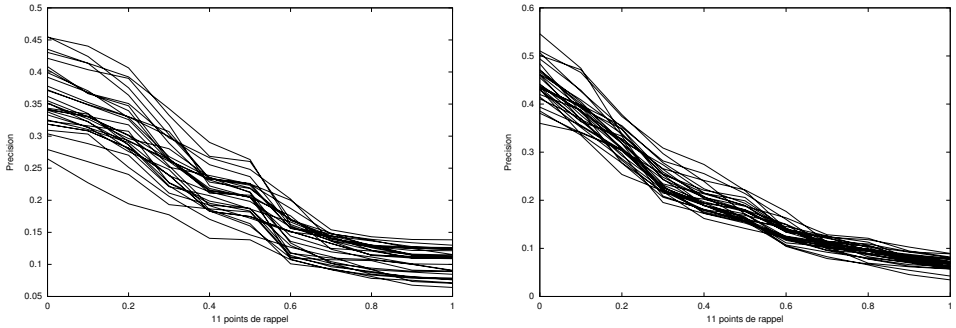
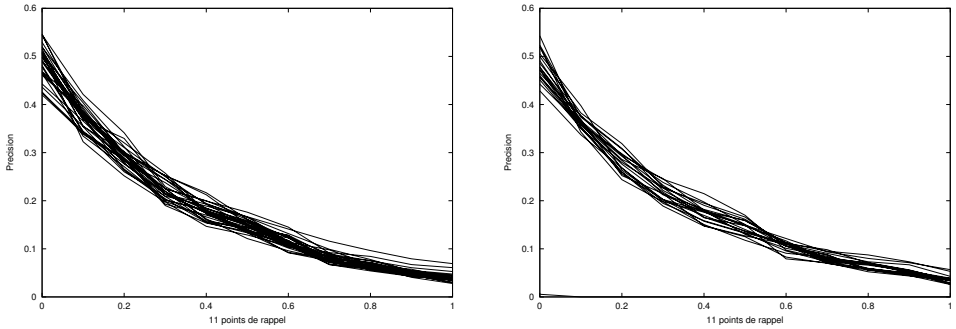


FIG. 4.17 – Rappel/précision pour les 30 sous-collections de WT10G à 600 000 documents et à 800 000 documents- modèle de *Clarke*



4.18, 4.19). Ceci semble logique : en effet, plus la taille de l'échantillon est petite, plus il est probable d'avoir des échantillons très différents (voire disjoints) ; *a contrario*, plus la taille de

FIG. 4.18 – Rappel/précision pour les 30 sous-collections de WT10G à 1 000 000 documents et à 1 200 000 documents- modèle de *Clarke*

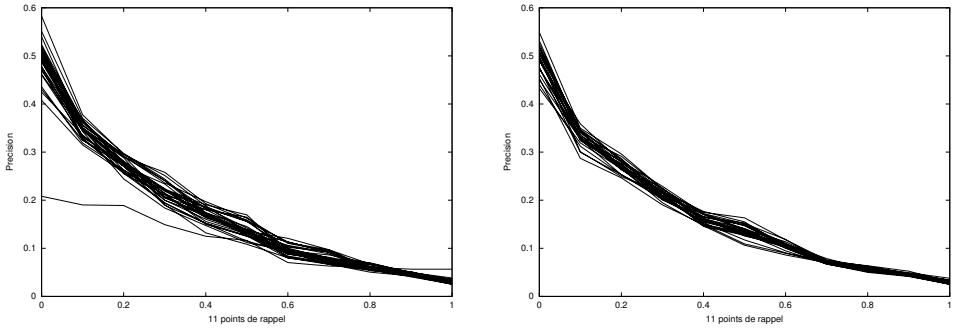
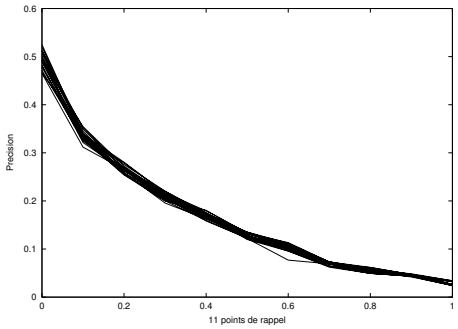


FIG. 4.19 – Rappel/précision pour les 30 sous-collections de WT10G à 1 400 000 documents- modèle de *Clarke*



l'échantillon est grande, plus elle se rapproche de celle de la taille de l'ensemble complet des documents et plus l'intersection entre les divers échantillons est grande (en d'autres termes plus les échantillons se ressemblent). Il n'est donc pas surprenant d'avoir des courbes très similaires (voire superposées) pour les échantillons de grandes tailles et des courbes plus différentes les unes des autres pour les échantillons de petite taille.
Pour les différents points standard de rappel, les moyennes et les médianes des précisions pour chaque taille de collection ont été observées. Les figures 4.20, 4.21, 4.22, 4.23 ,4.24 montrent les résultats.

Les courbes donnant les moyennes et les courbes donnant la médiane sont presque similaires. En effet, le constat est que de façon générale, la moyenne et la médiane de précision (à un niveau de rappel) sur les échantillons d'une taille sont proches. Ce qui montre que les valeurs de précision à un niveau de rappel sur les échantillons d'une taille ne sont pas très disparates. La construction des échantillons fait un contrôle sur la proportion de documents pertinents qu'elles contiennent. Le fait de changer le contenu des échantillons (choix aléatoire des documents) en

FIG. 4.20 – Rappel/précision : moyenne et médiane sur 30 sous-collections de chaque taille- Modèle de *Clarke*

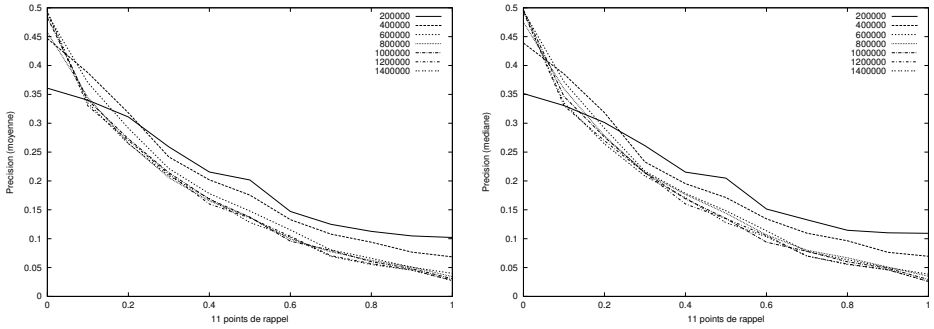


FIG. 4.21 – Rappel/précision : moyenne et médiane sur 30 sous-collections de chaque taille- Modèle de *Hawking*

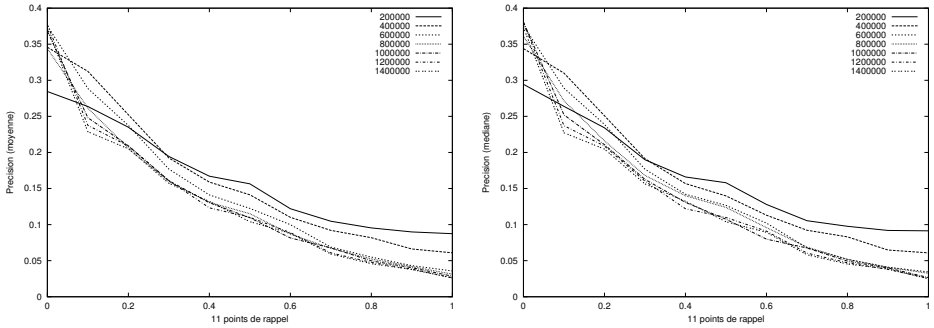
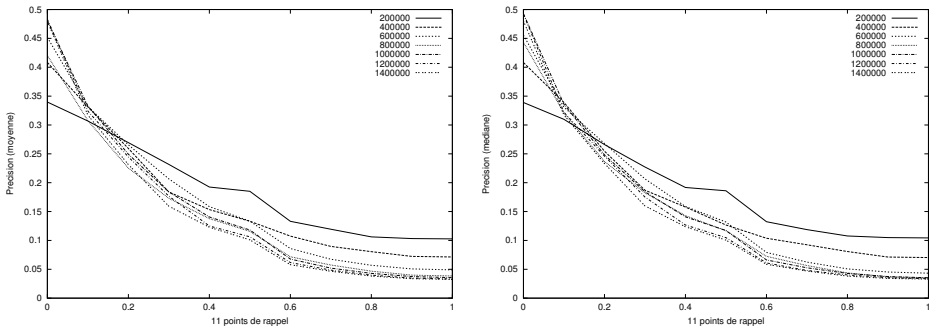


FIG. 4.22 – Rappel/précision : moyenne et médiane sur 30 sous-collections de chaque taille- Modèle de *Rasolofo*



gardant cette proportion affecte peu les performances des systèmes de RI utilisés, en ce qui concerne la performance rappel/précision.

L'impact de la taille sur les courbes rappel/précision est également analysé en construisant des boxplot sur la précision à chaque point de rappel et des intervalles de confiance sur la moyenne de ces précisions. Nous présentons les résultats pour les premiers niveaux de rappel.

FIG. 4.23 – Rappel/précision : moyenne et médiane sur 30 sous-collections de chaque taille- Modèle *okapip*

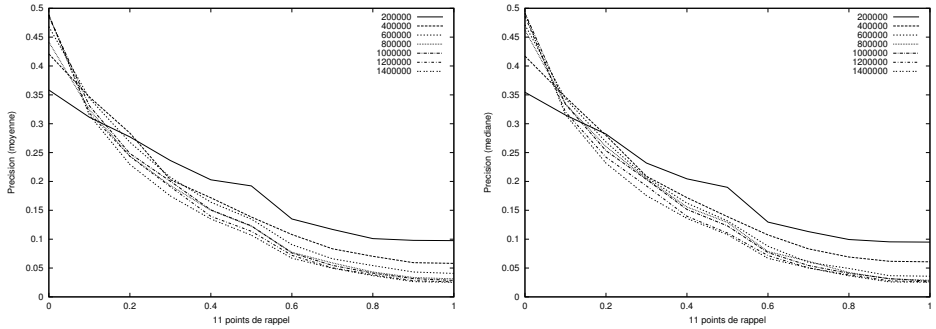
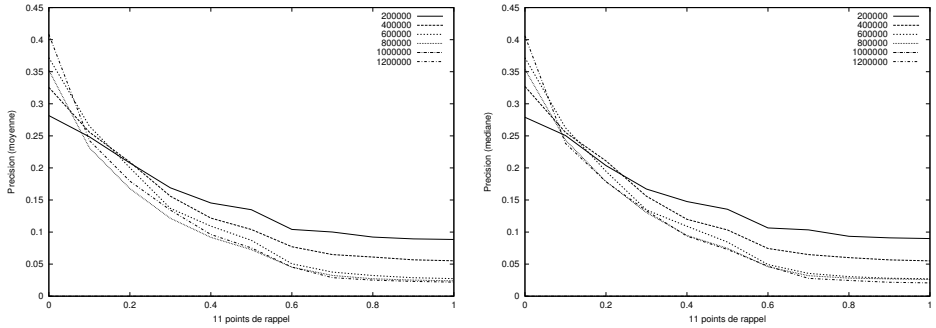


FIG. 4.24 – Rappel/précision : moyenne et médiane sur 30 sous-collections de chaque taille- Modèle de *MG*



Boxplot et intervalles de confiance sur la performance rappel/précision

FIG. 4.25 – Box-plot et intervalle de confiance sur PR0.0 : construit à base des 30 sous-collections de chaque taille - Modèle de *Clarke*

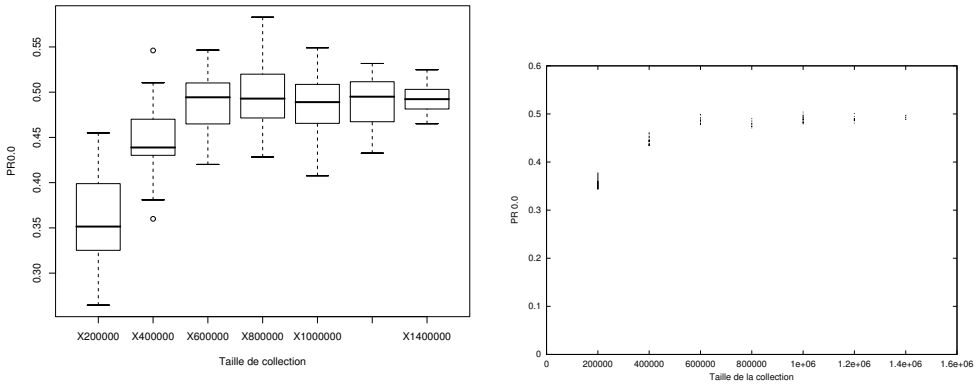


FIG. 4.26 – Box-plot et intervalle de confiance sur PR0.0 : construit à base des 30 sous-collections de chaque taille - Modèle de *Hawking*

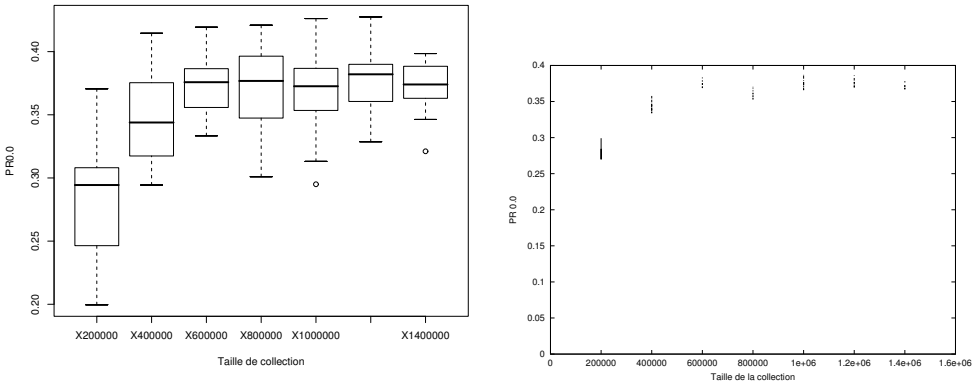


FIG. 4.27 – Box-plot et intervalle de confiance sur PR0.0 : construit à base des 30 sous-collections de chaque taille - Modèle de *Rasolofo*

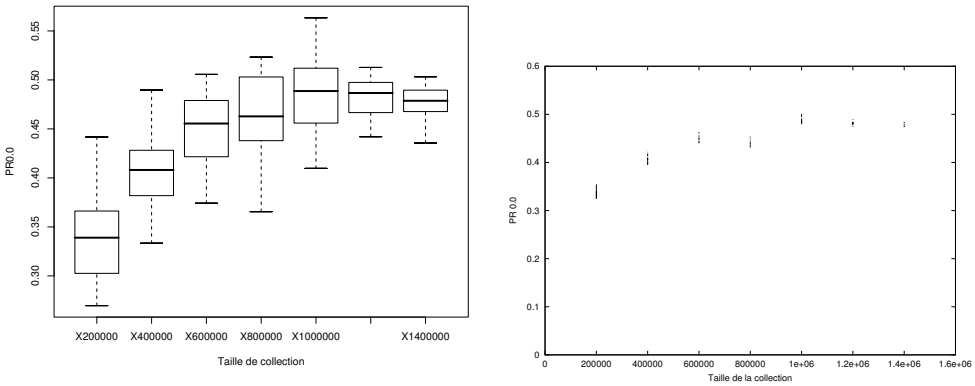


FIG. 4.28 – Box-plot et intervalle de confiance sur PR0.0 : construit à base des 30 sous-collections de chaque taille - Modèle vectoriel de *MG*

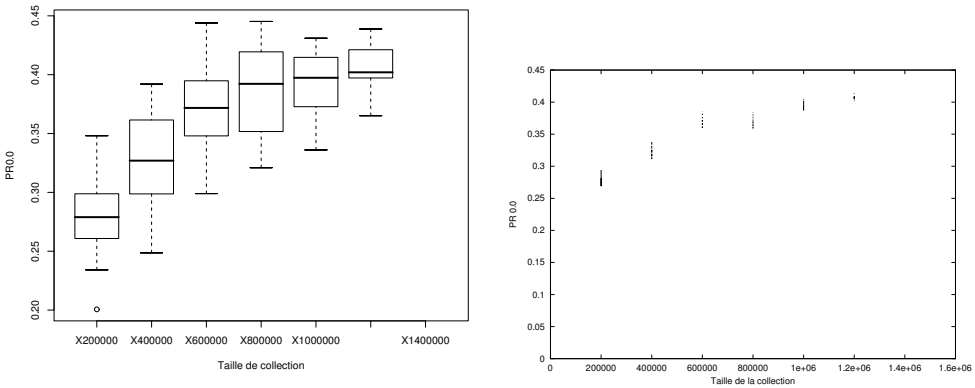


FIG. 4.29 – Box-plot et intervalle de confiance sur PR0.0 : construit à base des 30 sous-collections de chaque taille - Modèle Okapi

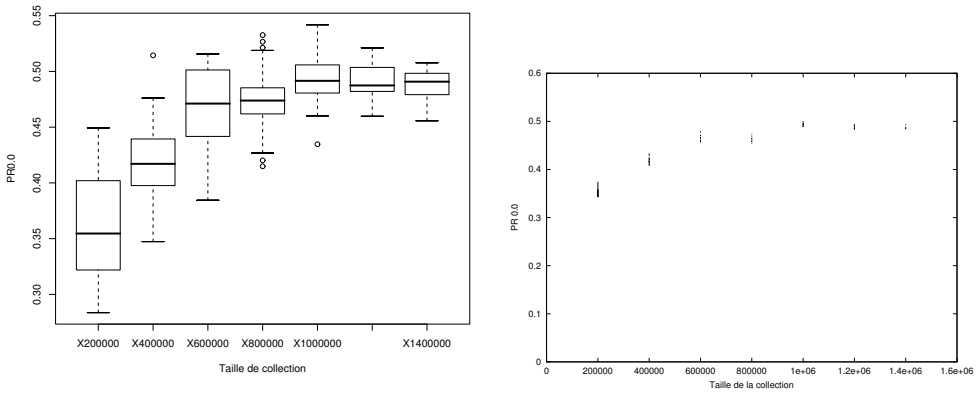


FIG. 4.30 – Box-plot et intervalle de confiance sur PR0.1 : construit à base des 30 sous-collections de chaque taille - Modèle de *Clarke*

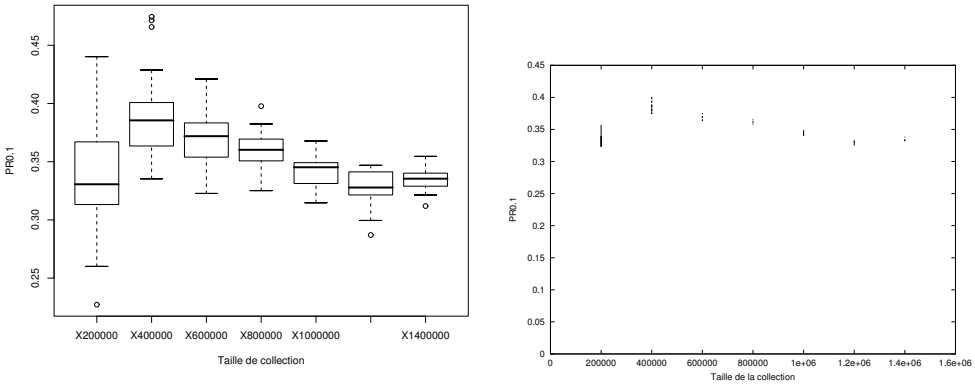


FIG. 4.31 – Box-plot et intervalle de confiance sur PR0.1 : construit à base des 30 sous-collections de chaque taille - Modèle de *Hawking*

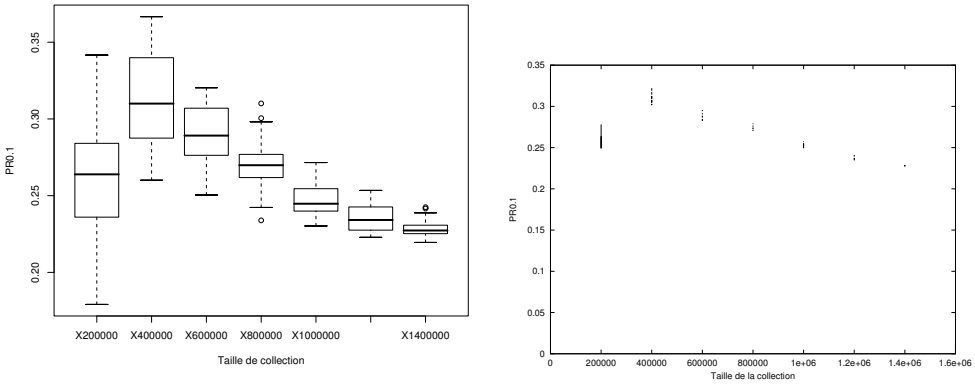


FIG. 4.32 – Box-plot et intervalle de confiance sur PR0.1 : construit à base des 30 sous-collections de chaque taille - Modèle de *Rasolofo*

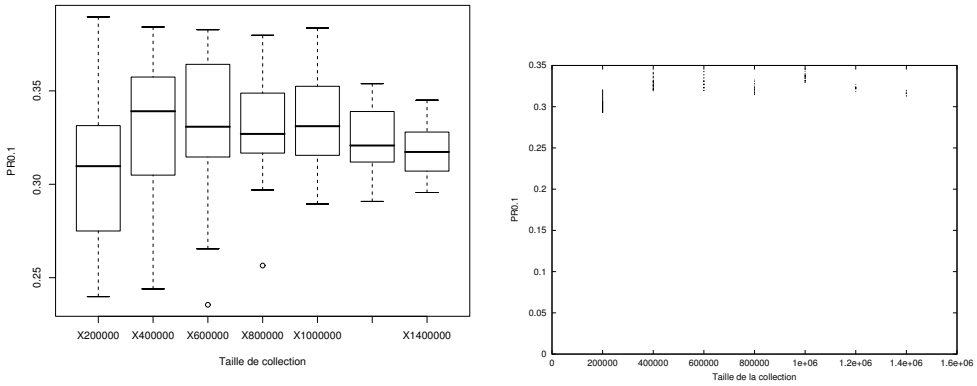


FIG. 4.33 – Box-plot et intervalle de confiance sur PR0.1 : construit à base des 30 sous-collections de chaque taille - Modèle vectoriel de *MG*

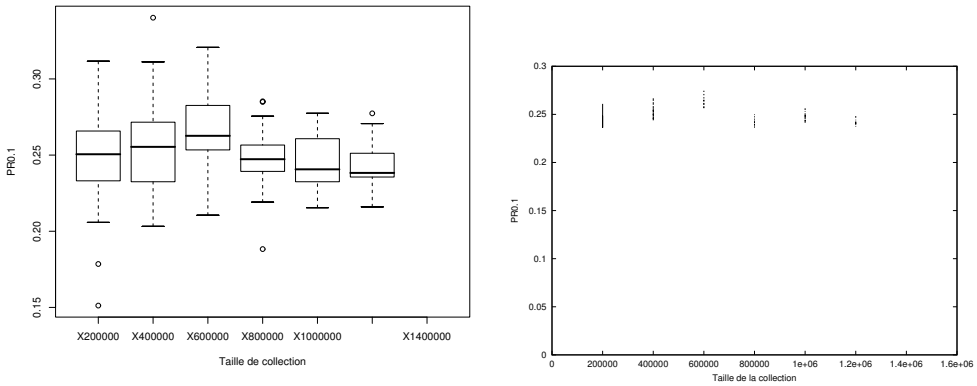
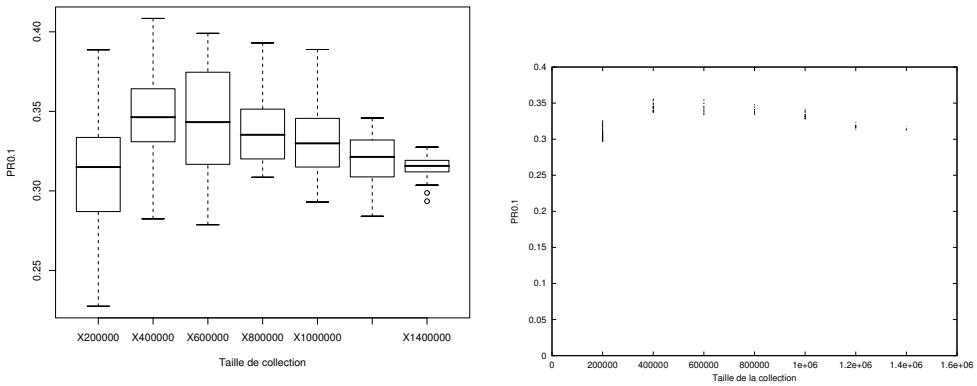


FIG. 4.34 – Box-plot et intervalle de confiance sur PR0.1 : construit à base des 30 sous-collections de chaque taille - Modèle Okapi

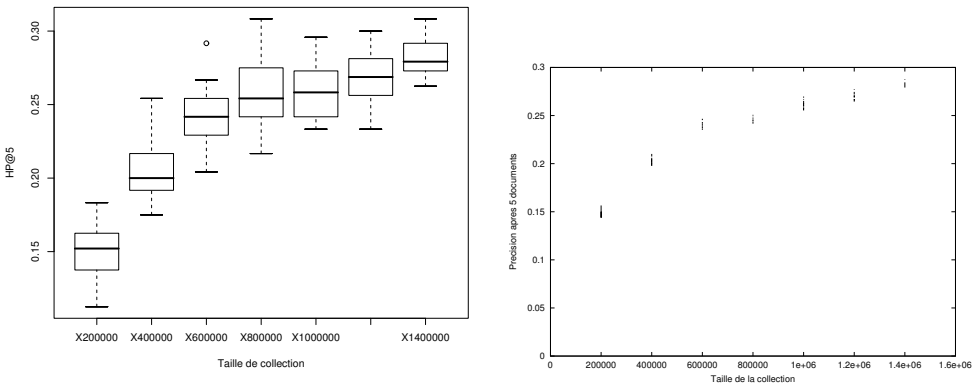


Pour les méthodes de *Clarke* et de *Hawking*, une croissance légère en début de croissance des collections est remarquée en ce qui concerne la précision au niveau de rappel 0.0. Pour les plus grandes collections, une constance de la précision au niveau de rappel 0.0, ce qui ne va pas dans le sens des résultats obtenus avec l'unique collection uniforme (première méthodologie).
Pour les méthodes de *Rasolofo* et *MG*, on note une légère croissance puis une baisse : pas de tendance ferme sur la précision au niveau de rappel 0.0. Pour le modèle *Okapip*, la croissance est plus marquée.
Des observations ont été faites sur la précision au niveau de rappel 10%. Les petites collections sont mieux positionnées pour les modèles de *Clarke* et de *Hawking*. Pour les méthodes de *Rasolofo*, *Okapip* et *MG*, les pertes de précision avec la croissance des collections est moins marquée, on a plutôt une certaine constance pour ce niveau de rappel.

Les boxplots et les intervalles de confiance renforcent l'analyse faite à partir des courbes de moyennes et de médiane sur les performances rappel/précision. Sur les tous premiers niveaux de rappel, les grandes collections ont des précisions meilleures (boxplot et intervalles de confiance sur PR0.0). Cette tendance s'inverse rapidement (dès le niveaux de rappel 10% à 20%).

Courbes de haute précision

FIG. 4.35 – Box-plot et intervalle de confiance sur P@5 : construit à base des 30 sous-collections de chaque taille - Modèle de *Clarke*



Toutes les courbes (box-plot et intervalles de confiance) de précision après un nombre de documents fixés montrent que la précision augmente avec la taille de la collection, comme les travaux de *Hawking et Robertson* présentés au chapitre 3 le montraient. Ceci est vérifié pour tous les modèles de RI utilisés dans ces expérimentations. Les courbes présentées concernent la précision après 5 documents retournés mais cette amélioration est constatée à divers points

FIG. 4.36 – Box-plot et intervalle de confiance sur P@5 : construit à base des 30 sous-collections de chaque taille - Modèle de *Hawking*

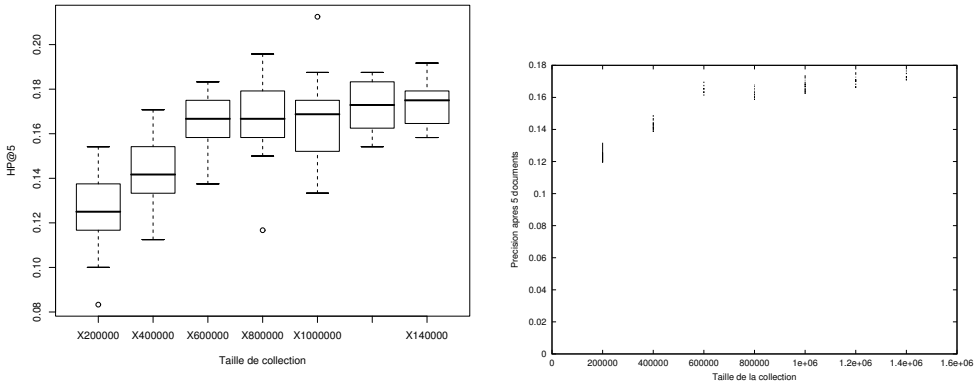
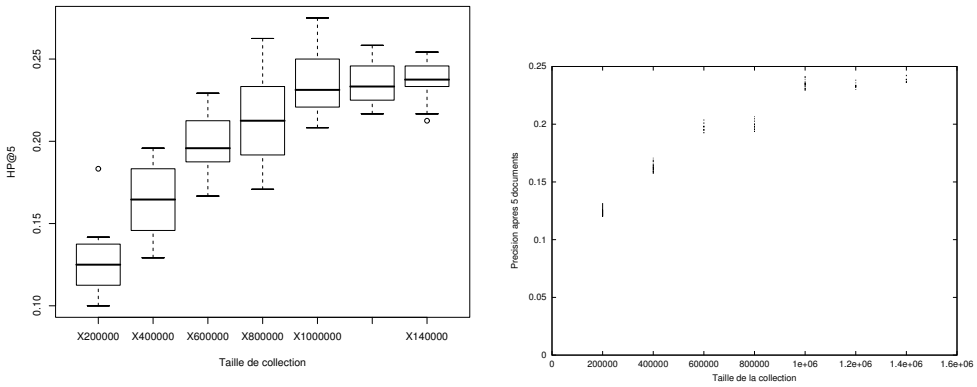


FIG. 4.37 – Box-plot et intervalle de confiance sur P@5 : construit à base des 30 sous-collections de chaque taille - Modèle de *Rasolofo*



de coupure (P@10, P@15, P@20, P@30). Comme l'ont suggéré ces auteurs, les raisons de cette croissance (inattendue) peuvent être associées au nombre de documents pertinents présents dans chaque sous-collection. Dans notre cas, ce nombre augmente avec la taille de collection (la proportion est la même donc la quantité augmente) et les SRI ont donc accès à plus de documents pertinents. Un SRI qui « sait trouver »les documents pertinents en tête de sa liste de résultats devrait donc améliorer sa haute précision. Toutefois, cet argument peut également être utilisé dans le sens contraire. En effet, pour les grandes collections, le SRI fait également face à une plus grande quantité de documents non pertinents.

FIG. 4.38 – Box-plot et intervalle de confiance sur P@5 : construit à base des 30 sous-collections de chaque taille - Modèle vectoriel de *MG*

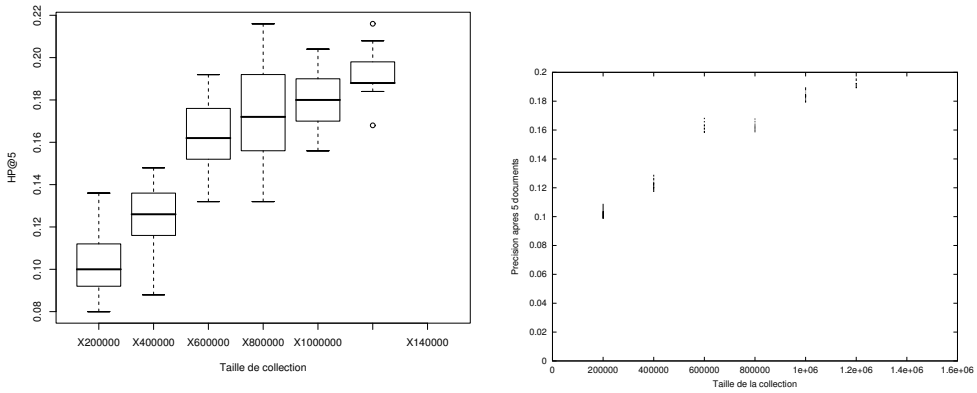
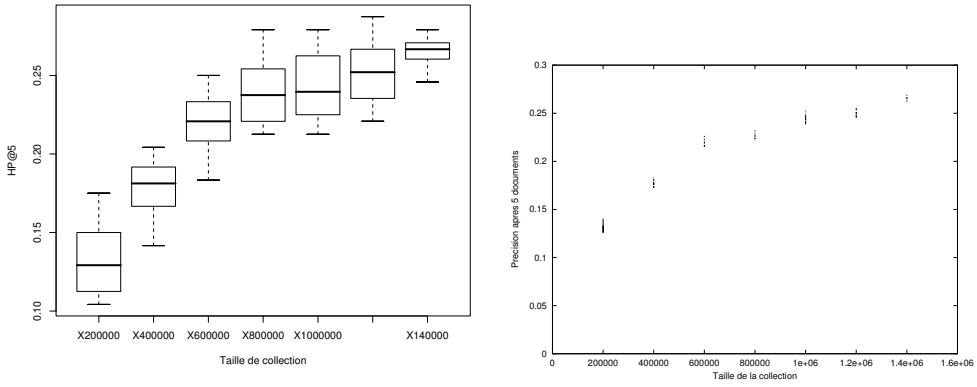


FIG. 4.39 – Box-plot et intervalle de confiance sur P@5 : construit à base des 30 sous-collections de chaque taille - Modèle Okapi



courbes de MAP

La précision moyenne (*Mean Average Precision*) est donnée par les figures 4.40, 4.41, 4.42, 4.44, et 4.43. On constate une baisse de cette précision moyenne avec l'accroissement des collections.

FIG. 4.40 – Box-plot et intervalle de confiance sur MAP : construit à base des 30 sous-collections de chaque taille - Modèle de *Clarke*

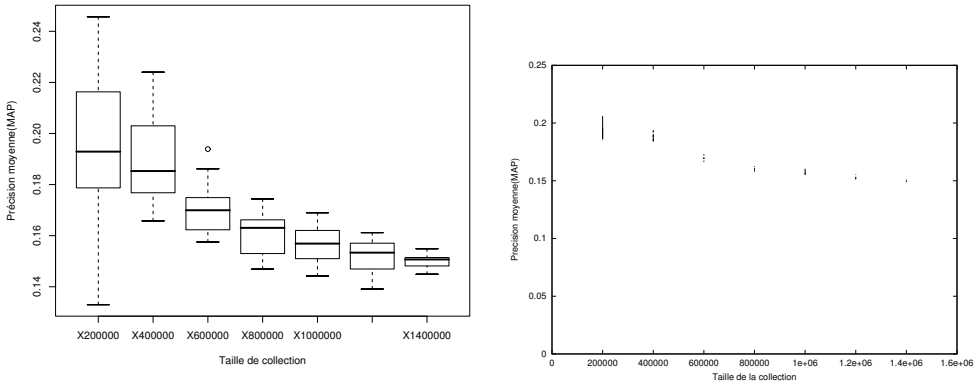
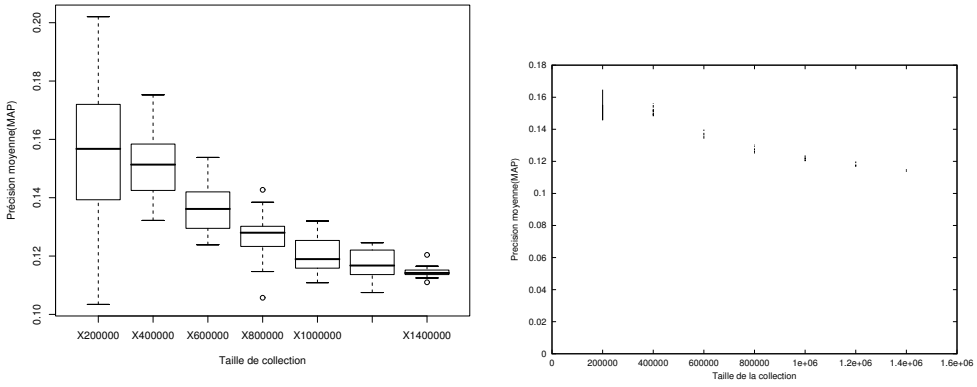


FIG. 4.41 – Box-plot et intervalle de confiance sur MAP : construit à base des 30 sous-collections de chaque taille - Modèle de *Hawking*



Courbes de R-Précision

La métrique R-précision est observée par les figures 4.45, 4.46, 4.47, 4.49, et 4.48. La moyenne et la médiane de R-précision est relativement stable avec la croissance des collections (une légère amélioration est notée). Sur les petites collections, la distribution de cette R-précision est assez étendue alors qu'elle est plus ressérée sur les grandes collections.

FIG. 4.42 – Box-plot et intervalle de confiance sur MAP : construit à base des 30 sous-collections de chaque taille - Modèle de *Rasolofo*

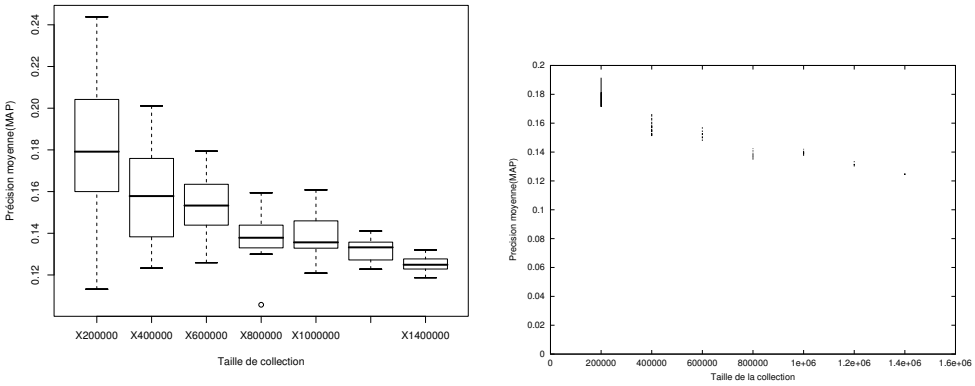


FIG. 4.43 – Box-plot et intervalle de confiance sur MAP : construit à base des 30 sous-collections de chaque taille - Modèle vectoriel de *MG*

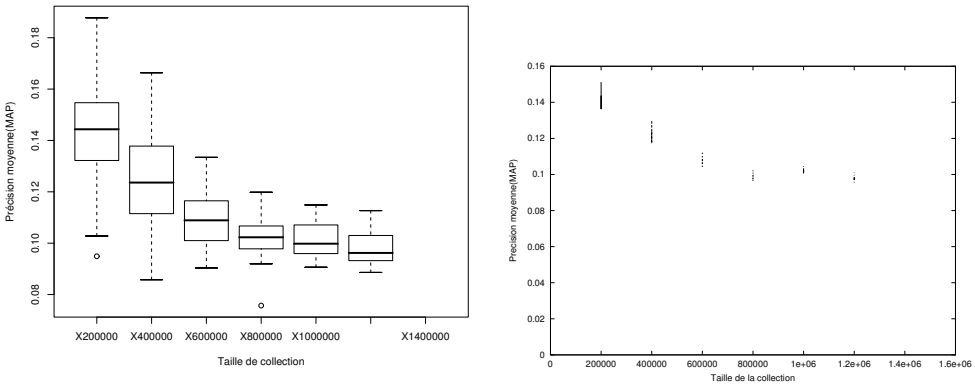


FIG. 4.44 – Box-plot et intervalle de confiance sur MAP : construit à base des 30 sous-collections de chaque taille - Modèle Okapi

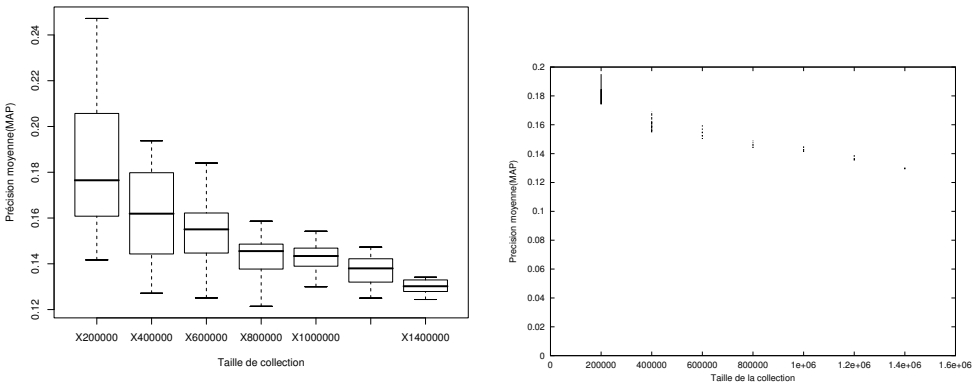


FIG. 4.45 – Box-plot et intervalle de confiance sur la R-précision : construit à base des 30 sous-collections de chaque taille - Modèle de *Clarke*

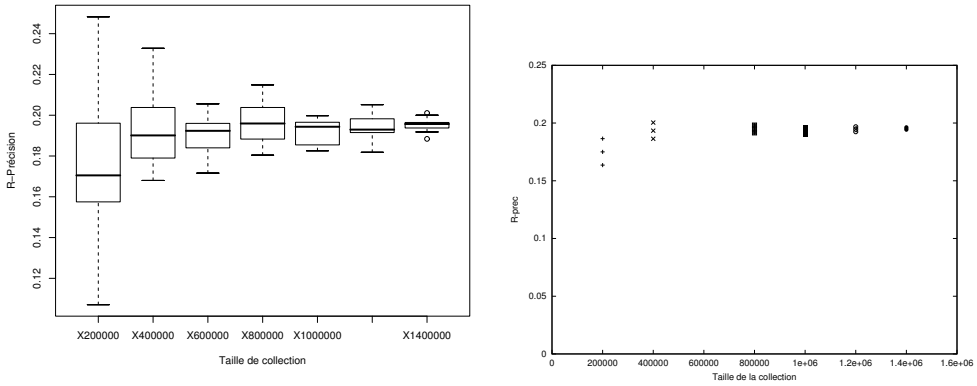
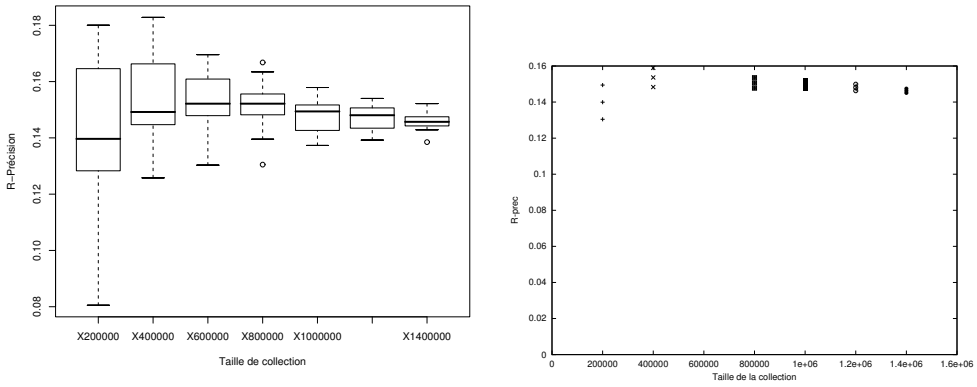


FIG. 4.46 – Box-plot et intervalle de confiance sur la R-précision : construit à base des 30 sous-collections de chaque taille - Modèle de *Hawking*



Apports

Cette seconde méthodologie de construction de collections de taille croissante a l'avantage de s'appuyer sur des techniques statistiques. Comme la première, elle affranchit les expérimentations sur l'impact du passage à l'échelle du biais lié au contenu des collections pour ne s'intéresser qu'au seul paramètre taille. L'utilisation d'une technique Monte-Carlo permet ensuite la construction de box-plots sur les distributions de performance et d'intervalles de confiance sur la moyenne de ces performances pour différentes sous-collections d'une même taille, ce qui permet d'interpréter les résultats obtenus de façon plus fiable sur le plan statistique.

Elle nous a permis de conforter certains résultats obtenus avec l'application de la première

FIG. 4.47 – Box-plot et intervalle de confiance sur la R-précision : construit à base des 30 sous-collections de chaque taille - Modèle de *Rasolofo*

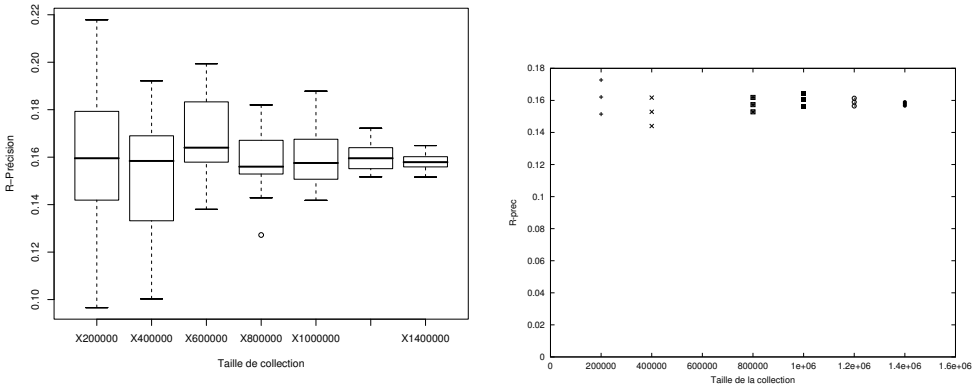
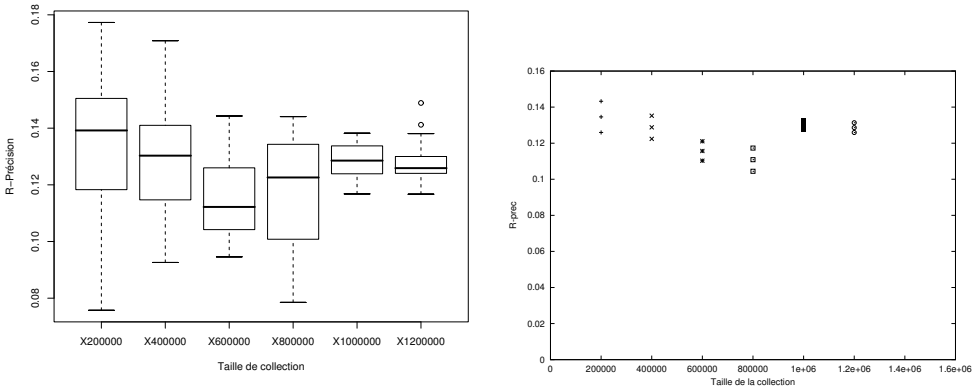


FIG. 4.48 – Box-plot et intervalle de confiance sur la R-précision : construit à base des 30 sous-collections de chaque taille - Modèle vectoriel de *MG*

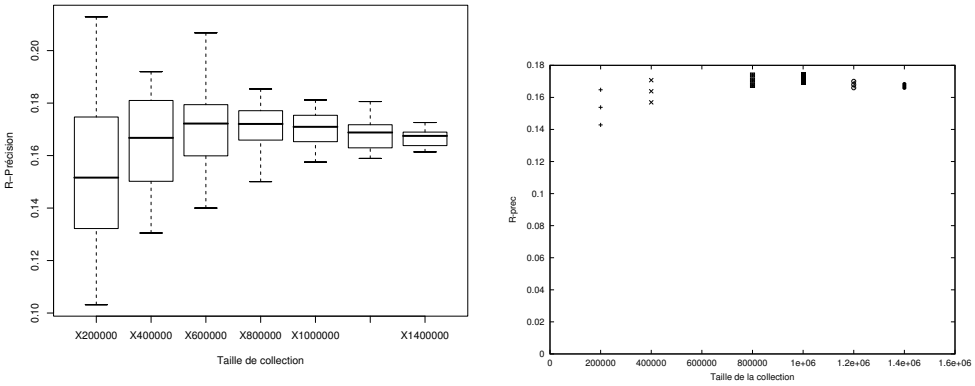


méthodologie (amélioration de la haute précision avec la taille de collection) mais elle nous a également apporté des résultats plus fermes sur l'impact de la taille sur d'autres métriques comme la MAP.

4.4 Synthèse

Dans ce quatrième chapitre, nous avons expliqué les méthodologies que nous proposons pour la construction de collections de taille croissante sur lesquelles il est possible de mener des expérimentations concernant le passage à l'échelle. La première permet de mener des expérimentations de façon aisément reproductible. La seconde découle de la première et en est une généralisation

FIG. 4.49 – Box-plot et intervalle de confiance sur la R-précision : construit à base des 30 sous-collections de chaque taille - Modèle Okapi



qui s'appuie sur la méthode statistique Monte-Carlo et permet ainsi une interprétation des résultats basés sur des techniques statistiques bien connues. Ces méthodes visent à affranchir les expérimentations sur le passage à l'échelle du biais lié au contenu des collections quand celles-ci augmentent en taille. Ainsi, le paramètre taille est le principal paramètre qui change au fil de la croissance et l'on peut donc observer son impact.

Ces méthodologies ont été appliquées à l'évaluation en RI et nous avons observé l'impact de la taille sur les performances des systèmes de RI (ces performances étant observées en utilisant des métriques classiques de la RI). Ces méthodologies peuvent être utilisées pour d'autres cas d'étude, en fonction des propriétés de systèmes de RI à étudier.

Ce chapitre a présenté le premier axe de nos travaux. Les métriques d'évaluation utilisées comme mesure de performance des systèmes de RI sont des métriques qui s'appuient sur une pertinence binaire. Dans le second axe, nous nous sommes intéressés à la pertinence multivaluée. Le chapitre 3 de cette thèse a présenté des travaux antérieurs qui s'y sont intéressés, notamment en proposant des métriques à pertinence multivaluée. Notre constat est cependant que les métriques existantes (pertinence binaire ou multivaluée) ne tiennent pas compte de façon directe du passage à l'échelle. Nous proposons des métriques qui visent directement à évaluer la capacité des systèmes de RI à améliorer leurs performances au cours du passage à l'échelle. Le prochain chapitre présente en détails ces travaux.

Chapitre 5

Evaluation du passage à l'échelle dans des environnements à pertinence multivaluée

Sommaire

Dans ce chapitre, nous présentons la partie de nos travaux qui porte sur des métriques pour le passage à l'échelle. L'impact de la croissance des collections sur les métriques existantes en RI peut être observé comme nous l'avons fait dans le premier axe de nos travaux, présenté au chapitre précédent. Cependant, ces métriques n'intègrent pas directement le passage à l'échelle en terme de taille des espaces de recherche. Dans cette partie, nous proposons des métriques dont le but est d'évaluer la capacité des systèmes de RI à passer à l'échelle : comme toutes les métriques centrées système, nos métriques s'appuient sur les documents pertinents ; elles prennent

en compte la pertinence multivaluée. Nous analysons donc les positions des documents retournés en fonction de leurs niveaux de pertinence au fil de la croissance des collections.

5.1 Introduction

La pertinence est le concept central de l'évaluation en RI. Le chapitre 2 a présenté ce concept socio-cognitif complexe sous divers angles et a montré sa place en RI. Elle peut être considérée comme une notion binaire ou multivaluée selon le niveau de détails souhaité et selon l'analyse qu'on souhaite en faire. Comme nous l'avons mentionné dans le chapitre 3, la pertinence comme notion binaire peut être restrictive à bien de points. Dans notre proposition de métriques, la pertinence multivaluée est prise en compte. Ces métriques sont orientées système et visent à évaluer la façon dont un système de RI donné passe à l'échelle. Nous observons les positions des documents retournés en fonction de leurs niveaux de pertinence au fil de la croissance des collections.
La première section de ce chapitre présente des concepts clés de nos métriques comme l'importance d'un degré de pertinence, le gain d'information réalisé en passant d'un degré de pertinence à un autre et nous réutilisons le principe du cumul de gain d'information par rang lorsqu'on parcourt la liste de résultats d'un système de RI (les travaux de *Kekäläinen et Järvelin* [80, 87, 88] ont introduit ce principe).

Les expérimentations menées en utilisant ces métriques sont analysées et une comparaison avec des métriques existantes qui utilisent la pertinence multivaluée est réalisée, notamment les métriques *Discounted Cumulative Gain* et *Cumulative Gain* de *Kekäläinen et Järvelin.*

La dernière section est une synthèse des éléments présentés dans le chapitre.

5.2 Métriques pour l'évaluation du passage à l'échelle

Les métriques que nous proposons introduisent deux concepts : l'importance d'un niveau de pertinence et le gain entre deux niveaux de pertinence. Dans les travaux antérieurs, le premier de ces concepts existe de façon tacite (travaux de *Kekäläinen et Järvelin* qui utilisent divers schémas de pondération pour *quantifier* les degrés de pertinence). Les propriétés (évidentes) liées à ce concept d'importance sont ici formalisées. Le second concept peut dans certains cas être défini en utilisant l'importance des degrés de pertinence et est différent du gain d'information cumulé utilisé par *Kekäläinen et Järvelin.* Cette différence est expliquée, avant d'introduire les métriques proprement dites.

5.2.1 Fonction d'importance d'un niveau de pertinence

L'hypothèse de départ est qu'à chaque document a été attribué un niveau de pertinence par rapport à chaque *topic*. Soit $\{niv_i\}$, $i = 1 \ldots, n$ l'ensemble des niveaux de pertinence de documents de la collection. Sur l'ensemble des documents on définit une relation d'équivalence R^t pour chaque *topic* t. Deux documents sont dans la même classe d'équivalence pour un *topic* donné s'ils ont le même niveau de pertinence au regard de ce *topic*.

Sur l'ensemble des niveaux de pertinence un ordre total que nous noterons $\succ$ est défini (tous les niveaux de pertinence sont comparables deux à deux). Ainsi dans les notations utilisées $niv_i \succ niv_j$ si $i > j$. Cette relation d'ordre total est suffisante pour donner la préférence souhaitée dans la liste des documents retournés mais elle ne donne pas d'indication sur l'importance à accorder à un niveau de pertinence par rapport aux autres (l'importance relative entre les niveaux de pertinence). Or c'est l'importance attribuée à un niveau de pertinence qui caractérise en fait la qualité/quantité d'information pertinente attendue d'un document ayant ce niveau de pertinence. On peut vouloir créditer (*resp.* pénaliser) fortement des systèmes qui retournent les documents ayant le plus haut niveau de pertinence en tête de liste (resp pas en tête de liste) : dans ce cas, ce plus haut niveau de pertinence doit avoir une importance élevée (par rapport aux autres niveaux de pertinence) lors de l'évaluation des SRI. Par exemple pour des applications où le but est de ne retenir que quelques documents mais de très hauts niveaux de pertinence. Toutefois, il existe aussi des applications pour lesquelles le but est d'avoir de nombreux documents de bon niveau de pertinence. Pour ce type d'applications, un SRI qui retourne des documents d'un bon niveau de pertinence et un SRI qui retourne des documents d'un très haut niveau de pertinence auront le même effet ; l'importance du « bon niveau »de pertinence et celle du « très haut niveau »de pertinence ne seront donc pas très éloignées.

Ainsi, une fonction I qui formalise l'importance des niveaux de pertinence dépendra de ce qu'on cherche à évaluer des SRI et du type d'applications sur lesquelles on voudrait pouvoir utiliser ces SRI. Toute fonction I caractérisant l'importance des niveaux de pertinence aura les propriétés suivantes (fonction positive et croissante) :

- $I(niv_i) > 0$ et $I(niv_i) > I(niv_j)$ si $niv_i \succ niv_j$ *i.e.* $i > j$

Des fonctions comme $I(niv_i) = a \times i + b$ ou $I(x) = a \times i^2 + b$ peuvent modéliser l'importance des niveaux de pertinence.

Exemples : On a 4 niveaux de pertinence « TP »pour Très Pertinent, « P »pour Pertinent , « FP »pour Faiblement pertinent, « NP »pour Non Pertinent, classés comme suit $TP \succ P \succ FP \succ NP$. On a les deux fonctions $I_1(niv_i) = i$ et $I_2(niv_i) = i^2$.

$I_1(TP) = 3$ et $I_1(FP) = 1$. $I_2(TP) = 9$ et $I_2(FP) = 1$. Pour la fonction I_1, TP est moins

« éloigné »de FP que si on choisit la fonction I_2.

Attribuer à chaque niveau de pertinence une valeur (qui caractérise son importance) dans l'absolu ne signifie rien, mais dans le relatif par rapport aux autres niveaux de pertinence cette valeur prend du sens.

5.2.2 Gain d'information entre deux niveaux de pertinence

Face à deux documents ayant des niveaux de pertinence différents pour un *topic*, on n'attend pas la même quantité d'information pertinente. Il est intéressant de pouvoir quantifier l'information pertinente gagnée (ou perdue) quand on passe d'un niveau de pertinence à un autre. Ce gain est une fonction des niveaux de pertinence : $Gain(niv_i, niv_j) = g(niv_i, niv_j)$. Quelles sont les caractéristiques d'une telle fonction g ?

- $g(niv_i, niv_j) > g(niv_i, niv_k)$ si $niv_j \succ niv_k$ *i.e.* si $j > k$ (1)
- $g(niv_i, niv_j) < g(niv_k, niv_j)$ si $niv_i \succ niv_k$ *i.e.* si $i > k$ (2)
- $g(niv_i, niv_i) = 0$ (3) : on ne gagne pas d'information pertinente quand on reste sur le même niveau de pertinence (même si on change de document, car les documents d'un même niveau sont de la même classe d'équivalence). On ne perd pas non plus d'information pertinente en restant sur le même niveau de pertinence.

Par déduction de (2) et (3), on a : $g(niv_i, niv_j) < 0$ si $niv_i \succ niv_j$ *i.e.* si $i > j$. En effet si on a $niv_i \succ niv_j$, alors cela signifie que la quantité d'information pertinente que contient tout document de niveau de pertinence niv_i est plus grande que la quantité d'information pertinente que contient tout document de niveau de pertinence niv_j. Donc, quand on passe d'un document de niveau de pertinence niv_i à un document de niveau de pertinence niv_j, on perd de l'information pertinente. De même, de (1) et (3) on déduit : $g(niv_i, niv_j) > 0$ si $niv_j \succ niv_i$ *i.e.* si $i < j$

La question du lien entre l'importance donnée aux niveaux de pertinence et le gain d'information réalisé entre deux niveaux de pertinence se posent. Vu la sémantique de ces deux concepts, il est logique que la fonction de gain d'information pertinente entre deux niveaux de pertinence dépende éventuellement de l'importance accordée à chacun des niveaux, *i.e.* de la valeur attribuée à chaque niveau de pertinence. Ainsi, l'on peut avoir $g(niv_i, niv_j) = fonction(I(niv_i), I(niv_j))$.

5.2.3 La distance mathématique comme exemple de fonction gain

De façon intuitive, le *gain* d'information pertinente entre deux niveaux de pertinence exprime une certaine notion de *distance* entre les deux niveaux de pertinence. Cette distance a les propriétés standards qui sont :

- (symétrie) : $\forall i, j, d(niv_i, niv_j) = d(niv_j, niv_i)$

- (séparation) : $\forall i,\ d(niv_i, niv_i) = 0$
- (inégalité triangulaire) : $\forall i,\ j,\ k\ d(niv_i, niv_j) \leq d(niv_i, niv_k + d(niv_k, niv_j)$

Puisque chaque niveau de pertinence est associé à une valeur numérique qui caractérise son importance, on peut construire la *distance* entre les différents niveaux de pertinence en se basant sur les valeurs numériques d'importance qui leur sont associées et
$d(niv_i, niv_j) = d(I(niv_i), I(niv_j))$. On remarque qu'en prenant :

$$\begin{cases} g(niv_i, niv_j) = -d(I(niv_i), I(niv_j)) \, pour \, niv_i \succ niv_j \\ g(niv_i, niv_j) = d(I(niv_i), I(niv_j)) \, pour \, niv_j \succ niv_i \end{cases}$$

on respecte bien toutes les propriétés attendues de la fonction g qui modélise le *gain* entre deux niveaux de pertinence. Ainsi, la notion de *distance* mathématique peut être utilisée pour mettre en œuvre notre fonction de *gain*.

Exemple : Nous reprenons l'exemple précédent de la section 5.2.1 et nous utilisons la distance mathématique de *Manhattan* donnant $d(x, y) =\mid x - y \mid$; on peut avoir $g(TP, FP) = -(4-2) = -2$ si on choisit I_1 comme fonction d'importance de niveau de pertinence et $g(TP, FP) = -(16-4) = -12$ si on choisit plutôt I_2.

5.2.4 Gain d'information cumulé à un rang donné

Des travaux antérieurs présentés dans le chapitre 3 ont proposé des métriques basées sur le gain d'information réalisé à chaque rang, lorsqu'on parcourt la liste des documents retournés par un système de RI sur une collection donnée. Dans les travaux de *Järvelin et Kekäläinen* [80], le gain d'information cumulé (*Cumulated Gain CG*) à un rang r est la somme cumulée des degrés de pertinence des documents de rang inférieur à r, soit en utilisant nos notations[46] :

$$\begin{cases} CG(1) = I(NivPertinence(d_1)) \\ CG(i) = CG(i-1) + I(NivPertinence(d_i)) \end{cases}$$

La métrique *Discounted Cumulative Gain (DCG)* calcule également des gains d'informations mais un cœfficient de pondération qui est une fonction décroissante du rang est appliqué.

$$\begin{cases} DCG(1) = I(NivPertinence(d_1)) \\ DCG(i) = DCG(i-1) + NivPertinence(d_i) \times^b \log(i) \quad \text{si} \quad i \neq 1 \end{cases}$$

A un rang donné r, l'on peut soit comparer le gain cumulé par rapport au gain cumulé idéal, soit visualiser les gains cumulés à chaque rang sous forme d'une courbe. La visualisation proposée pour nos métriques s'appuie sur ce même principe ; elle est expliquée en détail dans les sections à venir. Les métriques proposées par *Sakai* [125] sont basées sur la notion de gain cumulé.

[46] $NivPertinence(d)$ est le niveau de pertinence du document d comme nous l'introduisons dans la section 5.2.5

5.2.5 Cumuler les gains d'information pour évaluer

Le principe de cumul de gain d'information au fil des rangs a été réutilisé pour nos métriques. A un rang donné r les gains d'informations réalisés entre les documents-résultats de deux collections pour tous les rangs inférieurs à r sont cumulés. *Kekäläinen et Järvelin* cumulent plutôt ce que nous appelons l'importance des degrés de pertinence des documents-résultats d'un système de RI sur une collection. Deux catégories de métriques ont été proposées : les premières permettent d'évaluer l'évolution des performances d'un système de RI entre deux collections et les secondes permettent d'observer l'écart entre les performances d'un SRI sur une collection et les performances qu'aurait fourni le SRI *idéal*[47] sur la même collection de documents. Notre objectif est d'analyser ces écarts au fil de la croissance des collections et déterminer l'impact de la taille.

Evaluer entre deux collections : Métriques 1

Soient deux collections $C1$ et $C2$ de taille croissante et un SRI S. On souhaite analyser et évaluer le comportement de S sur chacune des deux collections ; notamment, on voudrait déterminer si les performances de S s'améliorent, restent stables ou se détériorent lorsque la taille de la collection augmente. Pour un *topic* donné t, à chaque rang, on va calculer la quantité d'information pertinente gagnée à ce rang lorsqu'on change de collection. Nos notations sont les suivantes :

- N est le point où l'on souhaite s'arrêter dans la liste des documents,
- Tout document d retourné a un niveau de pertinence qui sera noté $NivPertinence(d)$,
- Pour deux niveaux de pertinence niv_i et niv_j, on a le gain g qu'on réalise dans la qualité des résultats en passant du niveau de pertinence niv_i au niveau niv_j.

La fonction donnant à chaque rang k le document retourné est définie par :

$$d^t(C_i) : \begin{pmatrix} \mathcal{N} \rightarrow C_i \\ k \mapsto d^t_k(C_i) \end{pmatrix}$$

Puis nous notons $Retrieved^t(C_i) = d^t(C_i)(\mathbb{N})$: la liste de tous les documents retournés pour la collection C_i et $Retrieved^t_N(C_i) = d^t(C_i)([1, .., N])$: les N premiers documents retournés pour la collection C_i.

Pour simplifier les formules à venir, nous nommons *passage* entre les listes de résultats de deux collections C_i et C_j à un rang k le *gain* d'information pertinente entre le niveau de pertinence du document situé en position k de la liste de réponses de C_i pour le *topic* t et le niveau de

[47] Le SRI idéal retourne les documents par degré de pertinence. Nous avons réutilisé la conception de *Kekäläinen et Järvelin*

pertinence du document situé en position k de la liste de réponses de C_j pour le même *topic* t. On a :

$$Passage_k^t(C_i, C_j) = g(NivPertinence(d_k^t(C_i)), NivPertinence(d_k^t(C_j)))$$

Pour avoir le gain d'information pertinente relatif à un rang, un cœfficient de pondération (cp) qui dépend du rang est utilisé. Ce cœfficient de pondération (qui sera une fonction décroissante du rang) est nécessaire pour réduire progressivement en fonction du rang l'impact du gain/perte d'information pertinente. Comme suggéré par *Järvelin et Kekäläinen* [80], l'on peut faire cette réduction en pente rapide (avec une fonction comme l'inverse du rang $cp(k) = 1/k$) dans le cas où les documents en tête de liste sont ceux sur lesquels on veut focaliser l'évaluation ou en pente moins rapide (avec une fonction comme l'inverse du log du rang $cp(k) = 1/log_b(k)$).

Exemple tiré de [80] : Pour $b = 2$, $log_2(2) = 1$ et $log_2(1024) = 10$. Ainsi pour le rang 1024, on garderait quand même encore $1/10^{e}$du gain/perte d'information pertinente alors que avec une reduction en pente rapide avec par exemple $p(k) = 1/k$, on ne garderait que $1/1024^{e}$du gain/perte d'information pertinent pour le rang 1024.

Ainsi, pour chaque *topic* t, nous obtenons un vecteur de passages

$$< \ Passage_1^t(C_i, C_j), \ldots, Passage_N^t(C_i, C_j) \ >$$

et un vecteur de passages pondérés

$$< \ cp(1) \times Passage_1^t(C_i, C_j), \ldots, cp(N) \times Passage_N^t(C_i, C_j) \ >$$

Deux possibilités s'offrent à nous :

- La première est de calculer une somme des éléments du vecteur de passages pondérés d'un *topic* afin d'avoir une valeur unique pour chaque *topic* à une position donné N. Nous définissons donc la première métrique comme suit :

  $$Metrique1_N^t(C_i, C_j) = \sum_{k=1}^{N} cp(k) \times Passage_k^t(C_i, C_j)$$

  Une moyenne pourra ensuite être faite pour l'ensemble des *topics*. Cette métrique ne peut se calculer que sur l'ensemble de *topics* ayant des résultats sur chacune des deux collections C_i et C_j ; plus précisément sur des *topics* qui vérifient $\mid Retrieved^t(C_i) \mid \geq N$ et $\mid Retrieved^t(C_j) \mid \geq N$
- La deuxième possibilité est de calculer la somme des éléments des vecteurs (pondérés ou non) de tous les *topics* rang par rang, ce qui forunit un seul vecteur final de N éléments pour les *topics* :

  $$< cp(1) \times \sum_t (Passage_1^t(C_i, C_j)), \ldots, cp(N) \times \sum_t (Passage_N^t(C_i, C_j)) >$$

Ce vecteur-somme a la même taille que les vecteurs de passages de chaque *topic* ; son i^{e}élément est la somme des i^{e}éléments des vecteurs de chaque *topic*. On peut donc faire un cumul au fil des rangs et visualiser ces vecteurs sous forme de courbes gain/perte (sur tous les *topics*) cumulé par rang.

Evaluer par rapport au système dit idéal : Métriques 2

Le second groupe de métriques est basé sur le même principe que les Métriques 1. Pour une collection donnée C, le SRI S retourne une liste de résultats $Retrieved^{t}(C)$ pour un *topic* donné. La liste *idéale* de résultats pour ce *topic* est construite comme suit :

- soit $Documents^{t}(niv_i)$ l'ensemble des documents de niveau de pertinence niv_i pour le *topic* t et présents dans la collection C
- N est le point où l'on souhaite s'arrêter dans la liste des documents.

On crée une liste de documents $Retrieved\_ideal^{t}(C)$ qui est la liste « idéale »qui aurait pu être retournée par le SRI pour ce *topic*. Le principe est d'introduire dans cette liste d'abord les « meilleurs »documents, ensuite les bons et ainsi de suite. Pour chaque niveau de pertinence niv_i, en partant du plus pertinent vers le moins pertinent, on introduit nb documents « fictifs »ayant ce niveau de pertinence, avec :

$$nb = \begin{pmatrix} N - \sum_{j=i+1}^{n} \mid Documents(niv_j) \mid \text{ si } N - \sum_{j=i}^{n} \mid Documents(niv_j) \mid < 0 \\ sinon \mid Documents(niv_i) \mid \end{pmatrix}$$

Pour deux documents d_i et d_j pris dans cette liste, si d_i est placé avant d_j alors d_i est « plus pertinent »que d_j pour le *topic* ; autrement dit l'implication suivante est toujours vérifiée :

$$i < j \Longrightarrow NivPertinence(d_i) > NivPertinence(d_j)$$

Exemple : Nous reprenons l'exemple utilisé dans les sections 5.2.1 et 5.2.4, avec les niveaux de pertinence TP, P, FP et NP. Soit le *topic* t tel que $\mid Documents^{t}(TP) \mid = 7$, $\mid Documents^{t}(P) \mid = 10$, $\mid Documents^{t}(FP) \mid = 25$ et on choisit $N = 30$.

Pour chaque niveau de pertinence, on calcule le nombre de documents ayant ce niveau de pertinence à insérer dans la liste idéale de résultats de taille $N = 30$.

- Pour le niveau TP, on a $N - \sum_{j=3}^{3} \mid Documents(niv_j) \mid = 30 - 7 > 0$, on ajoute donc $\mid Documents(TP) \mid = 7$ documents de niveau de pertinence TP.
- Pour le niveau P, on a $N - \sum_{j=2}^{3} \mid Documents(niv_j) \mid = 30 - 7 - 10 > 0$, on ajoute donc $\mid Documents(niv_2) \mid = 10$ documents de niveau de pertinence P.

- Pour FP, on a $N - \sum_{j=1}^{3} \mid Documents(niv_j) \mid = 30 - 7 - 9 - 30 < 0$, on ajoute donc $N - \sum_{j=2}^{3} \mid Documents(niv_j) \mid = 30 - 7 - 10$ documents de niveau de pertinence FP.

La liste *idéale* formée sera constituée comme suit :

$$\underbrace{TP, \ldots, TP}_{7fois}, \underbrace{P, \ldots, P}_{10fois}, \underbrace{FP, \ldots, FP}_{13fois}$$

Comme dans le cas précédent concernant les Métriques 1, deux possibilités s'offrent à nous :

- soit nous calculons à un point donné N et pour un *topic* t,

  $Metrique2@N^t(C) = Metrique2_N^t(C) = Metrique1_N^t(Retrieved(C), Retrieved\_ideal^t(C))$

  Cette métrique donne le gain moyen sur la qualité qu'on réalise en passant de la liste de résultats obtenue sur la collection C par rapport à la liste *idéale* de résultats.
- soit nous calculons une somme des éléments des vecteurs pondérés de tous les *topics* rang par rang et on a un unique vecteur final de N éléments pour tous les *topics* :

  $$< cp(1) \times \sum_t (Passage_1^t(C, C\_ideal)), \ldots, cp(N) \times \sum_t (Passage_N(t)(C, C\_ideal)) >$$

  Ce vecteur somme a la même taille que les vecteurs de passages de chaque *topic*. Son $i$ème élément est la somme des $i$ème éléments des vecteurs de chaque *topic*. On peut ensuite faire un cumul au fil des rangs et visualiser les vecteurs sous forme de courbes gain/perte (sur tous les *topics*) cumulé par rang.

Nous constatons finalement que le concept qui va servir dans les métriques est celui de gain d'information entre deux niveaux de pertinence. Il peut être défini en s'appuyant sur le concept d'importance d'un niveau de pertinence mais finalement il peut aussi en être indépendamment défini. Dans les expérimentations que nous avons menées, nous explorons les deux cas de figure.

Lien avec les métriques existantes :

(Discounted) Cumulative Gain : En optant pour la distance mathématique de *Manhattan* $d(x, y) =\mid x - y \mid$ comme fonction de calcul des passages, nous retrouvons un lien avec les métriques *Cumulative Gain* et *Discounted Cumulative Gain*. En effet :

$DCG_N(C_i) - DCG_N(C_j)$

$$\begin{aligned}
&= \sum_{k=1}^{N} cp(k) I(NivPertinence(d_k^t(C_i))) - \sum_{k=1}^{N} cp(k) I(NivPertinence(d_k^t(C_j))) && (5.1)\\
&= \sum_{k=1}^{N} cp(k) [I(NivPertinence(d_k^t(C_i))) - I(NivPertinence(d_k^t(C_j)))] && (5.2)\\
&= \sum_{k=1}^{N} cp(k) g(NivPertinence(d_k^t(C_i)), NivPertinence(d_k^t(C_j))) && (5.3)\\
&= \sum_{k=1}^{N} Passage_k^t(C_i, C_j) && (5.4)\\
&= Metrique1_N^t(C_i, C_j) && (5.5)
\end{aligned}$$

De l'équation (5.2) à l'équation (5.3), nous utilisons le calcul de la fonction g décrit dans la section 5.2.3 utilisant une fonction distance. En effet, pour deux niveaux de pertinence niv_i et niv_j,

$$I(niv_i) - I(niv_j) = \begin{cases} d(I(niv_i), nivj) & si\ niv_i \succ niv_j \\ -d(I(niv_i), nivj) & si\ niv_i \prec niv_j \end{cases}$$

En posant de plus $cp(k) = 1\ \forall k$ (*i.e.* si on ne fait pas de pondération en fonction du rang), nous retrouvons cette même relation avec la métrique *Cumulative Gain*.
Le lien entre les métriques 2 et les métriques *Discounted Cumulative Gain* et *Cumulative Gain* se déduit directement du lien entre Métrique 1 et Métrique 2. Ainsi :
$DCG_N(C) - DCG_N(C_{ideal}) = Metrique2_N(C) = Metrique1_N(C, C_{ideal})$ si la fonction de calcul des passages choisie est la fonction distance mathématique de *Manhattan*.

Lien avec des métriques à pertinence binaire : Le cas de la pertinence binaire se modélise en attribuant des valeurs d'importance pris dans l'ensemble binaire {0, 1}. Dans ce contexte, il existe un lien entre les Métriques 1 et la précision après un nombre fixe de documents. En effet,

- si on choisit une fonction de gain entre niveaux de pertinence en utilisant la distance mathématique de *Manhattan*,
- si l'on choisit un cœfficient de pondération $cp(k) = 1\ \forall k$

alors $Metrique1^t_N(C_i, C_j)$ vaut $CG_N(C_i) - CG_N(C_j)$, soit $N \times (P@N(C_i) - P@N(C_j))$.

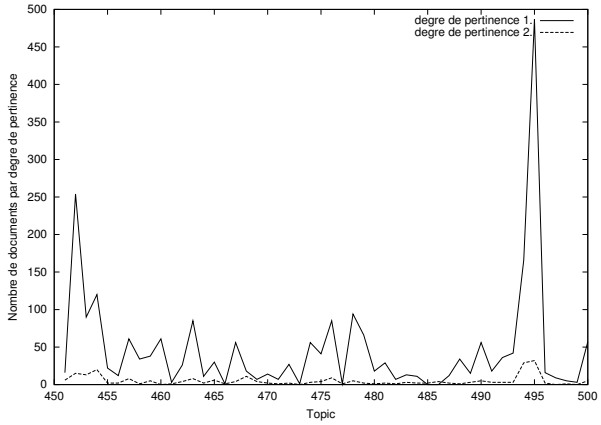
5.3 Expérimentations

5.3.1 Données

Collections, requêtes, modèles de RI

Dans le chapitre 4, la collection WT10G de TREC9 a été décrite, ainsi que les besoins d'informations liés à cette collection. Pour les expérimentations des métriques à pertinence multivaluée, cette même collection de test a été utilisée. Les jugements de pertinence fournis pour les *topics* 451-500 prennent en compte trois niveaux de pertinence (représentés par les valeurs 0, 1 et 2 dans les jugements fournis par *NIST*).
Les requêtes utilisées sont générées telles que décrites dans le chapitre 4. La figure 5.1 montre la répartition des documents par degré de pertinence pour chaque *topic*.
Les modèles de RI utilisés sont également ceux décrits dans le chapitre 4.

FIG. 5.1 – Nombre de documents par degré de pertinence et par *topic* dans WT10G, jugés par NIST (TREC9)



Protocoles d'expérimentation

Comme mentionné dans le chapitre 4, la répartition des documents pertinents (quel que soit le degré de pertinence) dans la collection WT10G est quelconque. La méthodologie de randomisation décrite dans le chapitre 4 en section 4.3 a été utilisée, pour générer des sous-collections *similaires* et de taille croissante et permettant de mettre en oeuvre les métriques proposées. La caractéristique de similarité des sous-collections est la proportion de documents par degré de pertinence. Ainsi, chaque sous-collection est construite en choisissant de façon aléatoire des documents, mais la proportion de documents par degré de pertinence pour chaque *topic* est identique à cette proportion dans la collection entière.

Pour chaque taille, 11 sous-collections ont été construites. Pour une taille donnée, les valeurs des métriques présentées sont des moyennes des valeurs obtenues pour chaque sous-collection de cette taille.

En ce qui concerne la fonction d'importance d'un degré de pertinence, le tableau 5.1 présente les schémas de pondération utilisés, inspirés de ceux de *Kekäläien* dans [88].

Pour le gain d'information entre deux niveaux de pertinence, dans un premier temps les distances mathématiques ont été utilisées (1er cas et 2nd cas). Dans un second temps, l'attribution des valeurs de gains d'information aux niveaux de pertinence se fait sans passer par une « fonction », modélisant ainsi le cas où le gain d'information entre deux niveaux de pertinence n'est pas linéaire (3e cas) et le cas où le gain d'information entre deux niveaux successifs n'est pas toujours le même (4e cas). Les tableaux 5.2, 5.3, 5.4 et 5.5 fournissent plus de détails.

Trois cœfficients de pondération ont été utilisés : $cp(k) = 1/k$, $cp(k) = 1/log_2(k)$, $cp(k) =$

TAB. 5.1 – Valeurs d'importance attribuées aux niveaux de pertinence (pour le calcul de DCG et CG notamment).

	schéma 1	schéma 2	schéma 3	
niv 3	2	10	100	(jugés par 2 dans WT10G)
niv 2	1	5	10	(jugés par 1 dans WT10G)
niv 1	0	0	0	(jugés par 0 ou non jugés dans WT10G)

TAB. 5.2 – Attribution des gains d'information entre niveaux de pertinence. 1erschéma : (2, 1, 0) - le gain d'information est calculé avec la distance de *Manhattan* entre les valeurs d'importance.

	niv 1	niv 2	niv 3
niv 1	0	1	2
niv 2	-1	0	1
niv 3	-2	-1	0

TAB. 5.3 – Attribution des gains d'information entre niveaux de pertinence. 2nd schéma : (10, 5, 0) - le gain d'information est calculé avec la distance de *Manhattan* entre les valeurs d'importance.

	niv 1	niv 2	niv 3
niv 1	0	5	10
niv 2	-5	0	5
niv 3	-10	-5	0

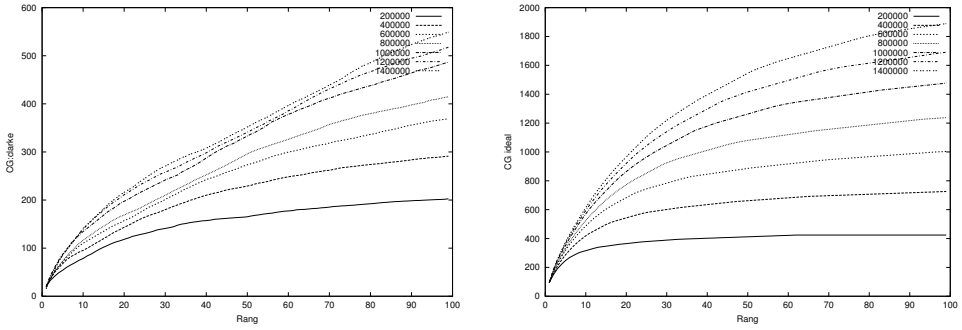
TAB. 5.4 – Attribution des gains d'information entre niveaux de pertinence. 3^{e}schéma : 100, 10, 0), le gain d'information est calculé avec la fonction $g(niv_i, niv_j) = signe(i-j) \times pas^{|i-j|}$ si $i \neq j$, 0 sinon. Le pas choisi vaut 10.

	niv 1	niv 2	niv 3
niv 1	0	10	100
niv 2	-10	0	10
niv 3	-100	-10	0

TAB. 5.5 – Attribution des gains d'information entre niveaux de pertinence (pas de correspondance avec les valeurs d'importance). Le gain d'information entre deux niveaux de pertinence successifs n'est pas toujours le même.

	niv 1	niv 2	niv 3
niv 1	0	10	100
niv 2	-10	0	50
niv 3	-50	-10	0

FIG. 5.2 – CG par taille de collection (modèle de *Clarke* et cas idéal) jusqu'au rang 100 - Schéma (0, 1, 2).



$1/log_{10}(k)$. Le premier cas modélise une réduction en pente rapide en fonction du rang (les premiers rangs sont donc très importants) alors que les deux derniers cas correspondent à une réduction plus lente (voir section 5.2.5).

5.3.2 Résultats

En plus des métriques que nous proposons, parmi les métriques qui utilisent également la notion de pertinence multivaluée, les métriques *Cumulative Gain* et *Discounted Cumulative Gain* ont été implémentées. Pour la métrique DCG, les cœfficients de pondération $cp(k) = 1/log_2(k)$ et $cp(k) = 1/log_{10}(k)$ ont été utilisés. Une version de la métrique DCG qui utilise plutôt le cœfficient de pondération $cp(k) = 1/k$ a également été mise en oeuvre. Cette version est propre à nos travaux et ne correspond pas à la définition de la métrique DCG telle que fournie par [87].

Description des courbes : 1erschéma de pondération (0, 1, 2)

Les courbes de CG et DCG sont données de la figure 5.2 à la figure 5.7. Les résultats fournis portent sur les rangs jusqu'à 100 et ensuite ils focalisent sur les premiers rangs.

FIG. 5.3 – CG par taille de collection (modèle de *Clarke* et cas idéal) Les 1ers rangs uniquement - Schéma (0, 1, 2).

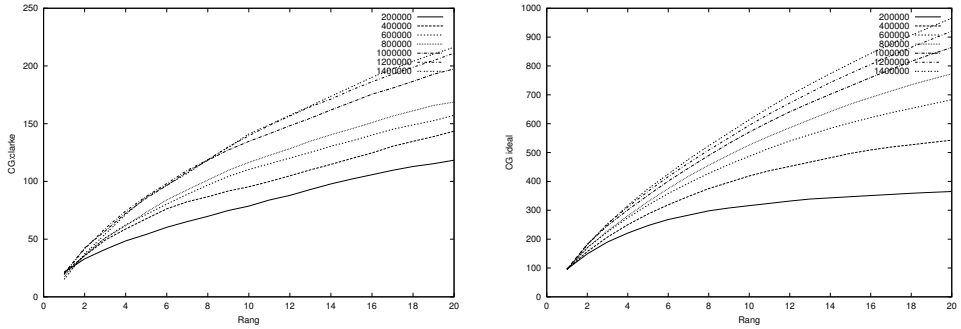


FIG. 5.4 – DCG avec cœfficient de pondération 1/log2(rang), par taille de collection (modèle de *Clarke* et cas idéal - jusqu'au rang 100 - Schéma (0, 1, 2).

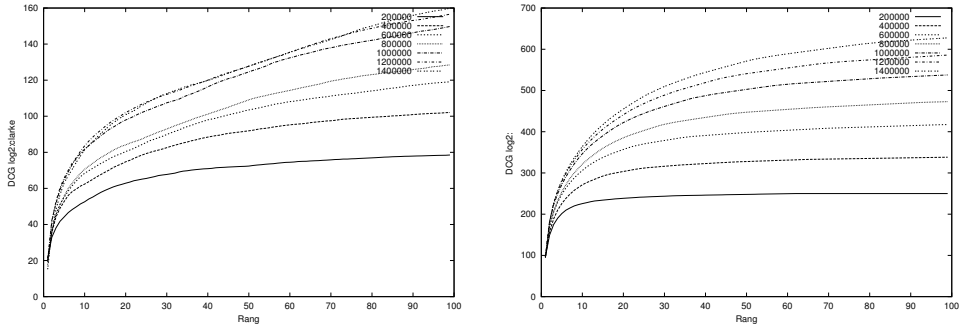


FIG. 5.5 – DCG avec cœfficient de pondération 1/log10(rang), par taille de collection (modèle de *Clarke* et cas idéal - jusqu'au rang 100 - Schéma (0, 1, 2).

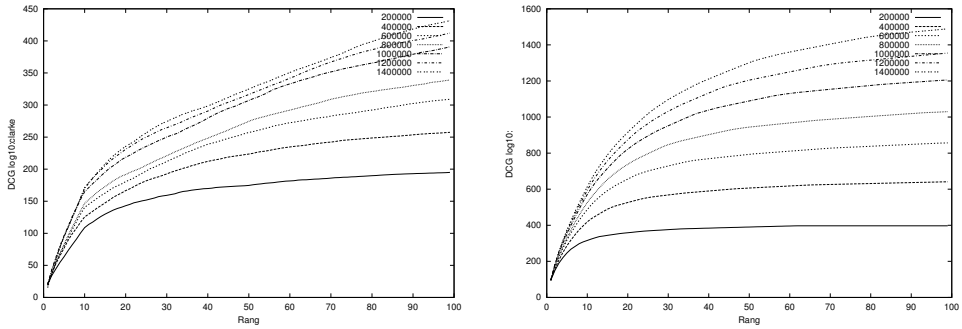


FIG. 5.6 – DCG avec cœfficient de pondération 1/log2(rang) par taille de collection (modèle de *Clarke* et cas idéal) Les 1ers rangs uniquement - Schéma (0, 1, 2).

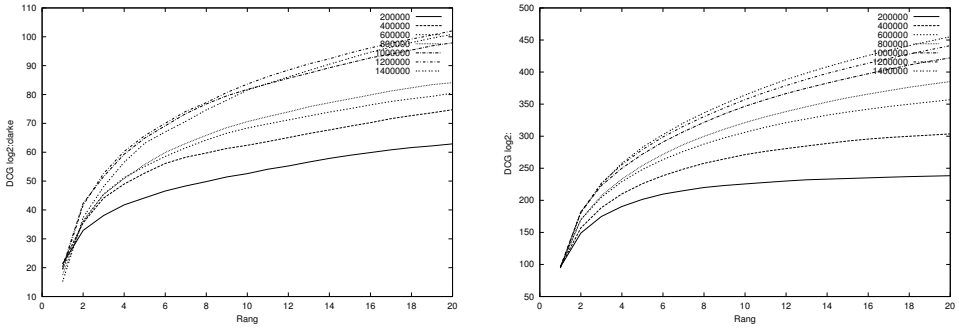


FIG. 5.7 – DCG avec cœfficient de pondération 1/log10(rang) par taille de collection (modèle de *Clarke* et cas idéal) Les 1ers rangs uniquement - Schéma (0, 1, 2).

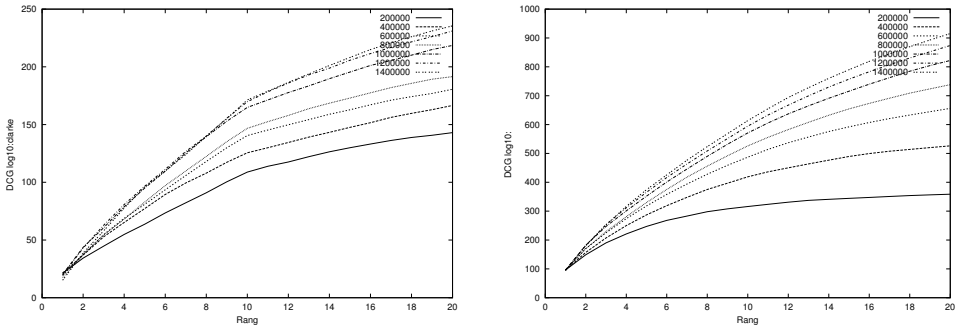


FIG. 5.8 – Métrique 1 - cumul par rang - cœfficient de pondération avec 1/log2(rang) - méthode de *Clarke* - Schéma (0, 1, 2).

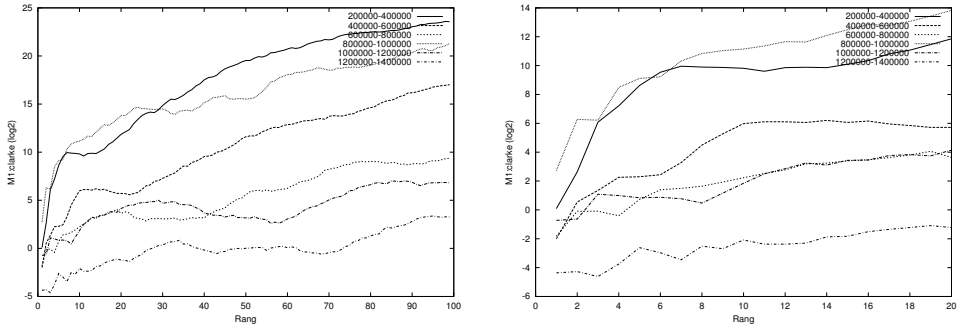
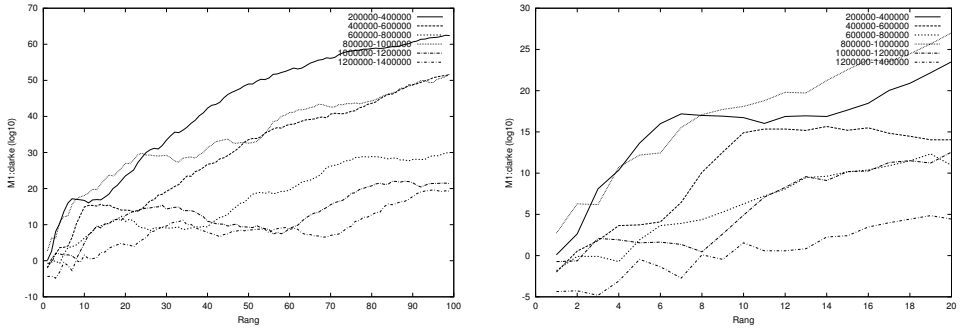


FIG. 5.9 – Métrique 1- cumul par rang - cœfficient de pondération avec 1/log10(rang) - méthode de *Clarke* - Schéma (0, 1, 2).



Les courbes portant sur les métriques Métriques 1 (cumul par rang) sont données par les figures 5.8 à 5.9, pour le premier schéma de pondération des niveaux de pertinence (0, 1, 2).

Les courbes portant sur les métriques Métriques 2 (cumul par rang) sont données par les figures 5.10 à 5.11, pour le premier schéma de pondération des niveaux de pertinence (0, 1, 2).

Les courbes portant sur les métriques Métriques 2 (calculées à un point fixe Métriques 2@N) ont également été fournies. La figure 5.12 et la figure 5.13 portent sur le modèle de *Clarke*, pour le premier schéma de pondération des niveaux de pertinence (0, 1, 2), au point de coupure $N = 5$, avec différents cœfficients de pondération. La figure 5.14 correspond au point de coupure $N = 10$ pour ce même modèle.

Pour les autres modèles, la figure 5.16, la figure 5.15 et la figure 5.17 fournissent les résultats pour le point de coupure $N = 5$.

Nous obtenons des courbes similaires pour les autres points de coupure (10, 20, 30).

FIG. 5.10 – Métrique 2 avec cœfficient de pondération 1/log2(rang)- cumul par rang - Méthode de *Clarke* - Schéma (0, 1, 2).

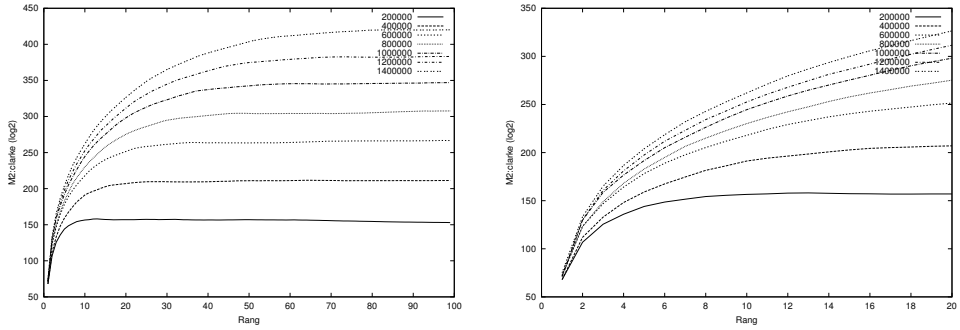


FIG. 5.11 – Métrique 2 avec cœfficient de pondération 1/log10(rang) - cumul par rang - Méthode de *Clarke* - Schéma (0, 1, 2).

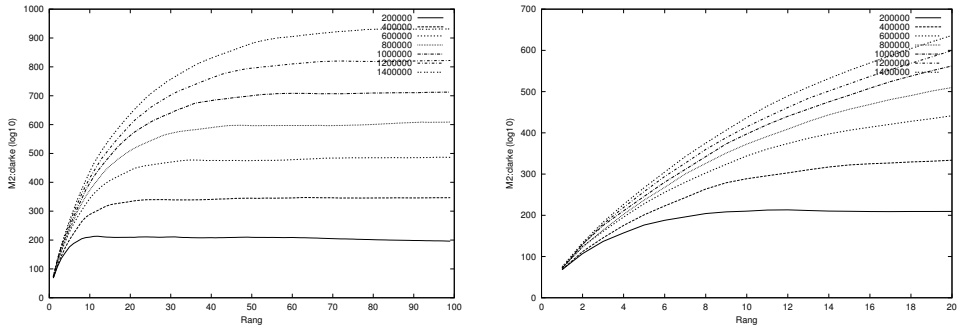


FIG. 5.12 – Métrique 2 sans cœfficient de pondération et avec le cœfficient de pondération 1/rang - cut-off 5 - Méthode de *Clarke* - Schéma (0, 1, 2).

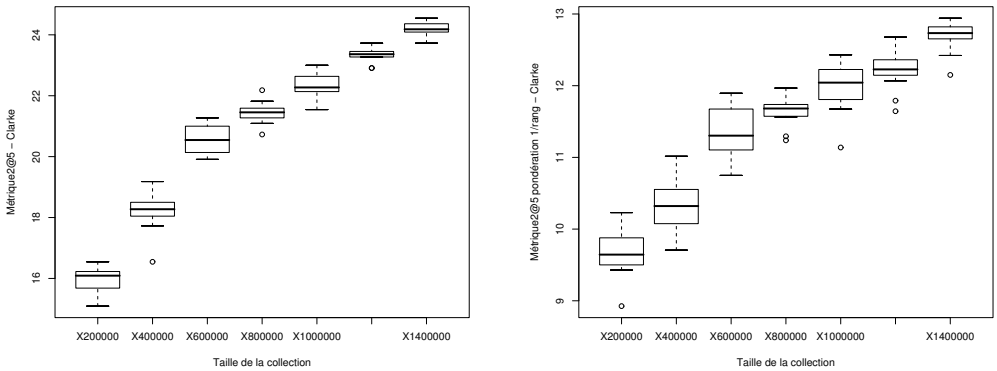


FIG. 5.13 – Métrique 2 avec cœfficient de pondération 1/log2(rang) et 1/log10(rang) - cut-off 5 - Méthode de *Clarke* - Schéma (0, 1, 2).

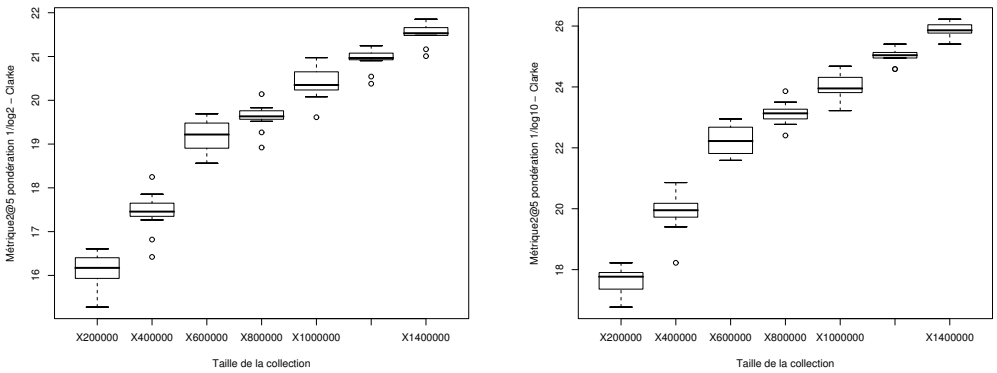


FIG. 5.14 – Métrique 2 avec cœfficient de pondération 1/log2(rang) et 1/log10(rang) - cut-off 10 - Méthode de *Clarke* - Schéma (0, 1, 2).

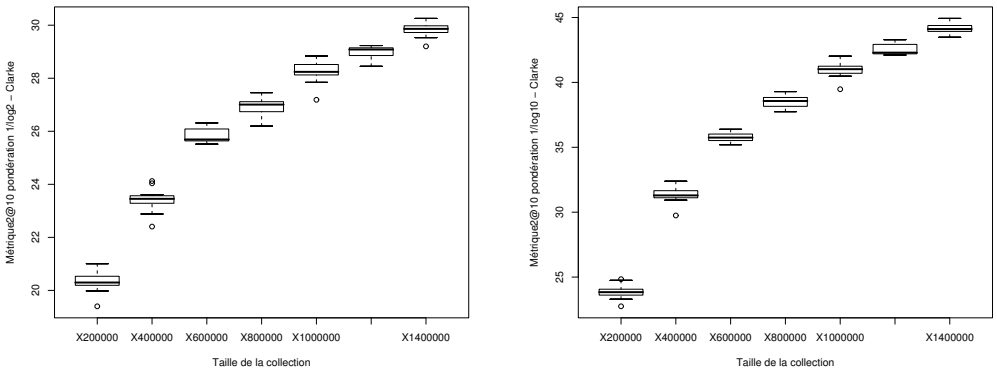


FIG. 5.15 – Métrique 2 avec cœfficient de pondération 1/log2(rang) - cut-off 5 - Modèle de *Lucy* et modèle *Okapip* - Schéma (0, 1, 2).

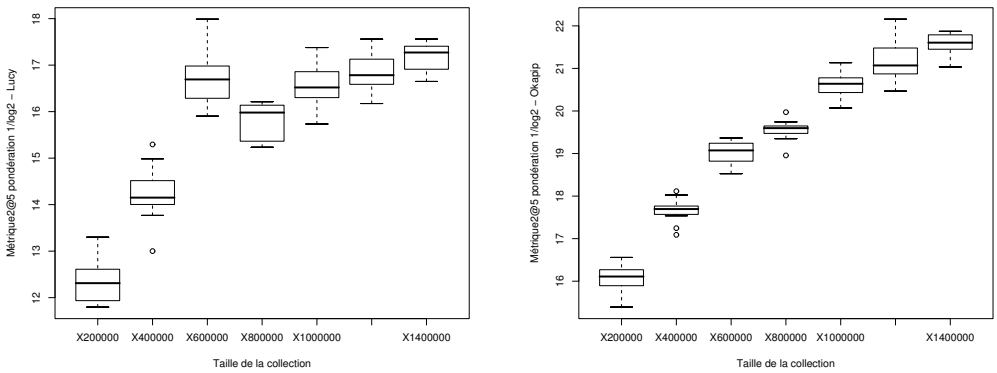


FIG. 5.16 – Métrique 2 avec cœfficient de pondération 1/log2(rang) - cut-off 5 - Modèle de *Hawking* et modèle de *Rasolofo* - Schéma (0, 1, 2).

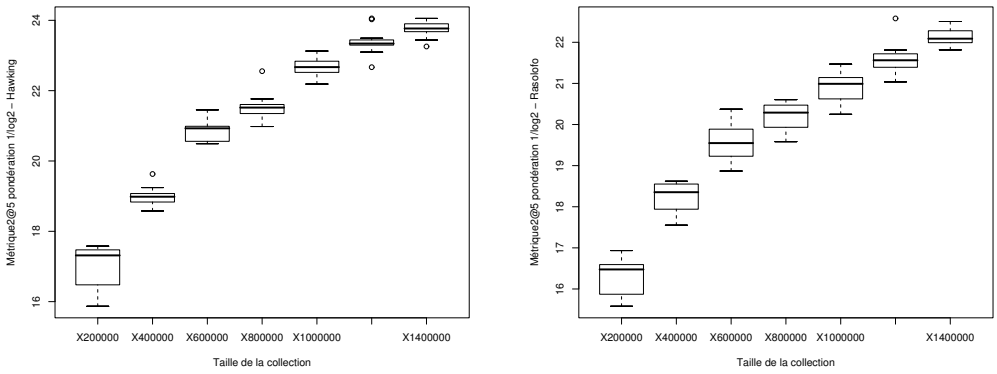


FIG. 5.17 – Métrique 2 avec cœfficient de pondération 1/log2(rang) - cut-off 5 - Modèle vectoriel de *MG* - Schéma (0, 1, 2).

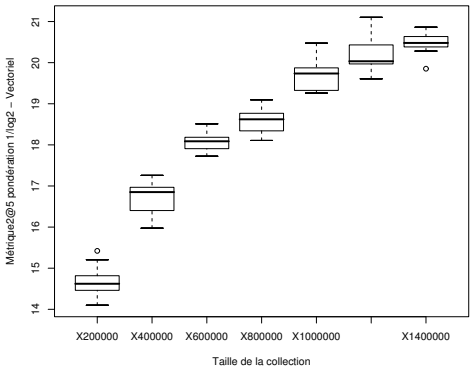


FIG. 5.18 – CG par taille de collection (modèle de *Clarke* et cas idéal) jusqu'au rang 100

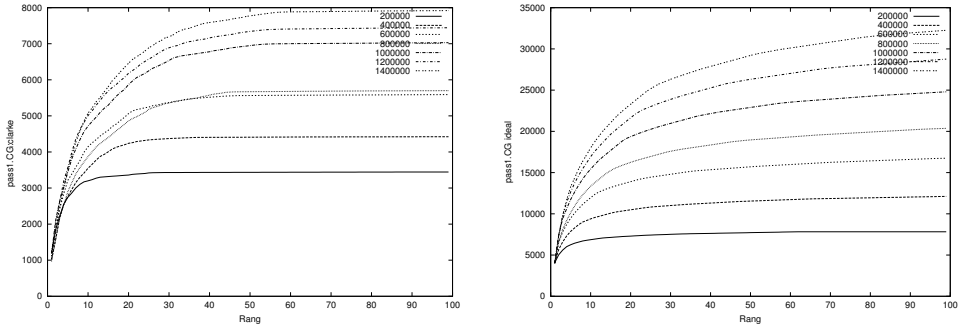
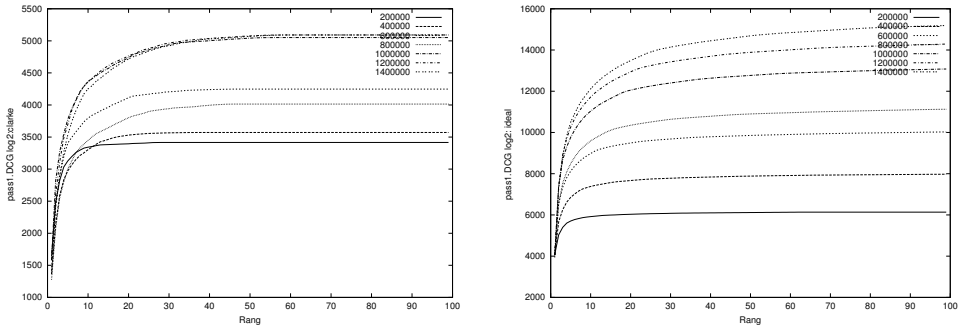


FIG. 5.19 – DCG avec 1/log2(rang), par taille de collection (modèle de *Clarke* et cas idéal) jusqu'au rang 100 - Schéma (0, 10, 100).



Description des courbes : 2nd schéma de cœfficient de pondération (0, 10, 100)

Les figures du second schéma de pondération vont de la figure 5.18 à la figure 5.23.

Description des courbes : 3e schéma de pondération (0, 10, 50, 100)

Les figures du troisième schéma de pondération vont de la figure 5.24 à la figure 5.27.

Description des courbes : 4e schéma de pondération (0, 5, 10)

Les figures du quatrième schéma de pondération vont de la figure 5.28 à la figure 5.34.

Interprétation

Le CG fournit le cumul de gain d'information au fil du parcours de la liste de résultats. Les courbes de CG par taille de collection montrent un positionnement de ces dernières par ordre suivant la taille de collection. A un rang donné, le CG augmente globalement avec la taille de la collection à partir des rangs autour de 5. Pour les tous premiers rangs, les courbes de CG se

FIG. 5.20 – Métrique 1 sans cœfficient de pondération et avec 1/rang - cumul par rang - méthode de *Clarke* - Schéma (0, 10, 100).

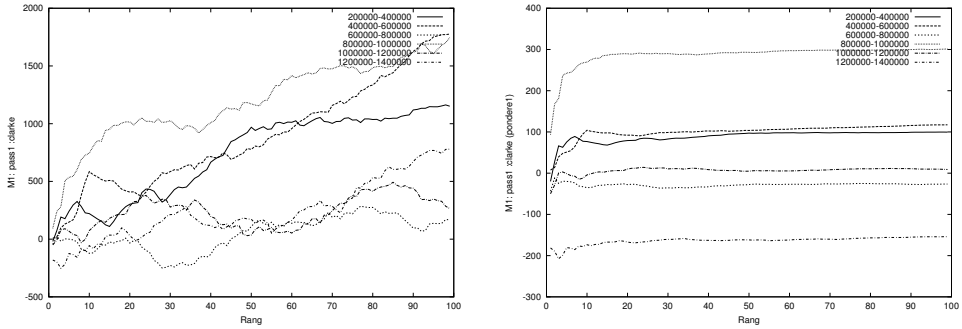


FIG. 5.21 – Métrique 1 avec le cœfficient de pondération et avec 1/log2(rang) - cumul par rang - méthode de *Clarke* - Schéma (0, 10, 100).

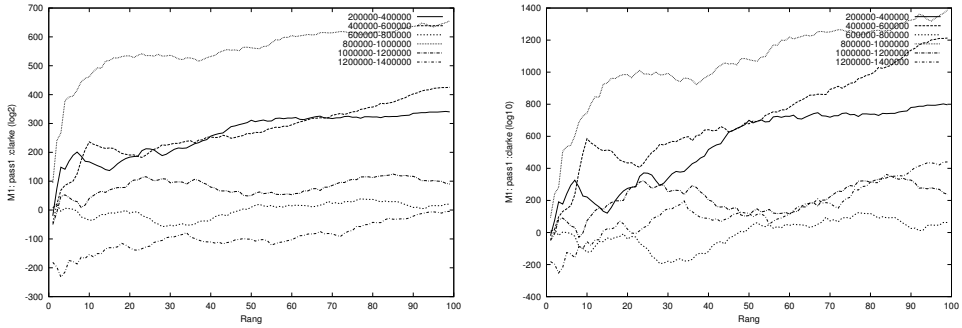


FIG. 5.22 – Métrique 2 sans cœfficient de pondération et avec 1/rang - cumul par rang - méthode de *Clarke* - Schéma (0, 10, 100).

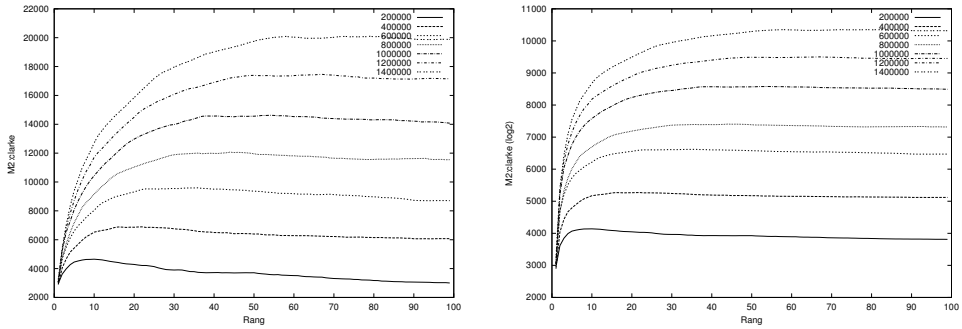


FIG. 5.23 – Métrique 2 avec les cœfficients de pondération 1/log2(rang) et avec 1/log10(rang) - cumul par rang - méthode de *Clarke* - Schéma (0, 10, 100).

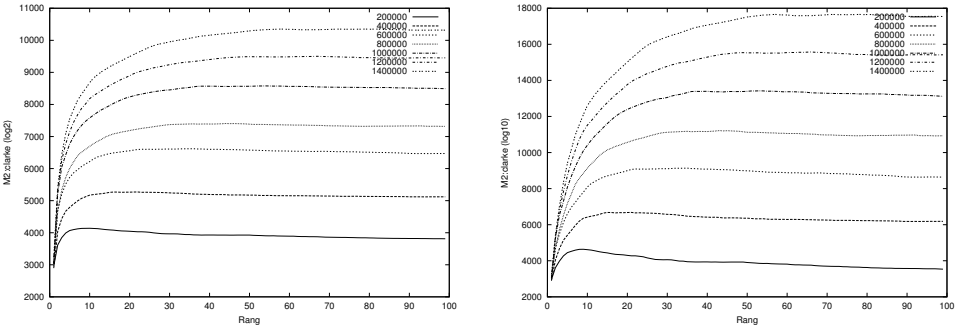


FIG. 5.24 – Métrique 1 sans cœfficient de pondération et avec 1/rang - cumul par rang - méthode de *Clarke* Schéma (0, 10, 50, 100).

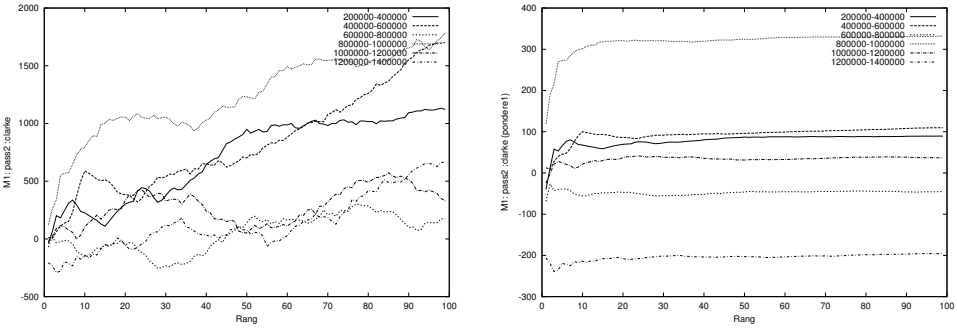


FIG. 5.25 – Métrique 1 avec les cœfficients de pondération 1/log2(rang) et 1/log10(rang) - cumul par rang - méthode de *Clarke* Schéma (0, 10, 50, 100).

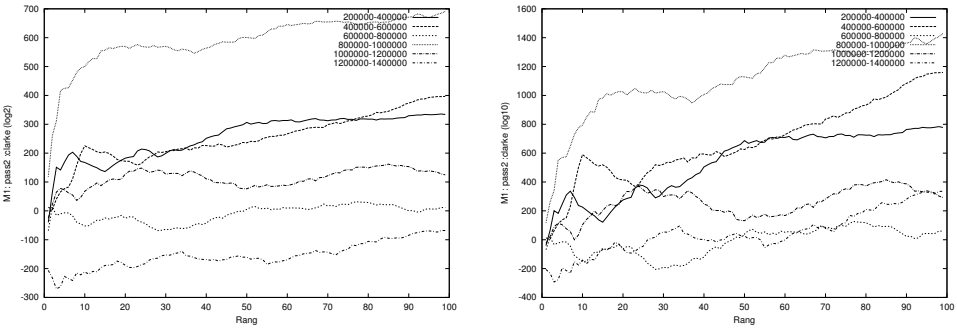


FIG. 5.26 – Métrique 2 sans cœfficient de pondération et avec 1/rang- cumul par rang - méthode de *Clarke* - Schéma (0, 10, 50, 100).

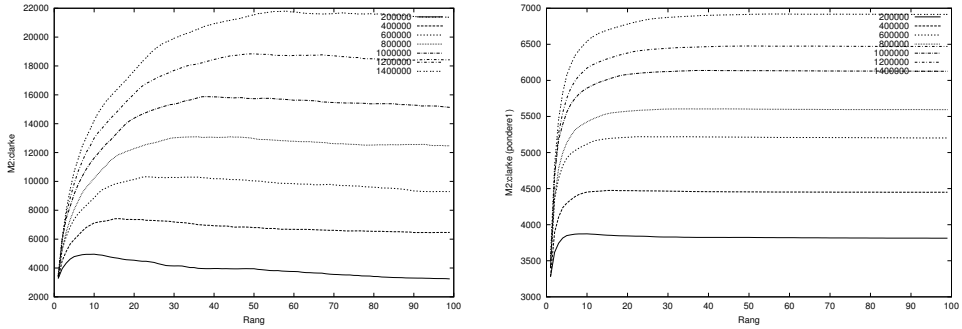


FIG. 5.27 – Métrique 2 avec les cœfficients de pondération 1/log2(rang) et 1/log10(rang) - cumul par rang - méthode de *Clarke* - Schéma (0, 10, 50, 100).

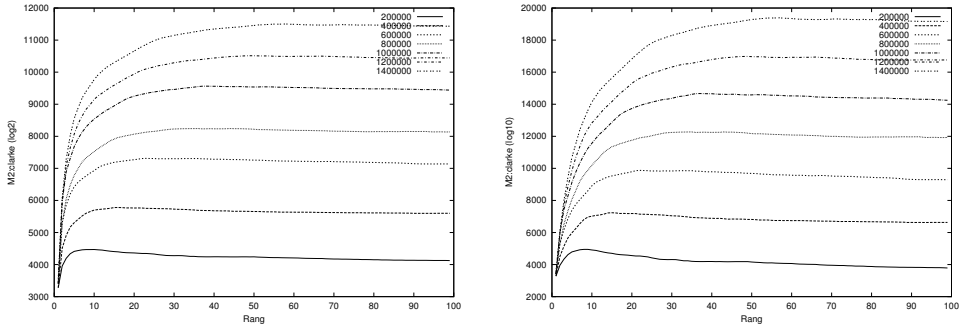


FIG. 5.28 – CG par taille de collection (modèle de *Clarke* et cas idéal) jusqu'au rang 100 Schéma (0, 5, 10).

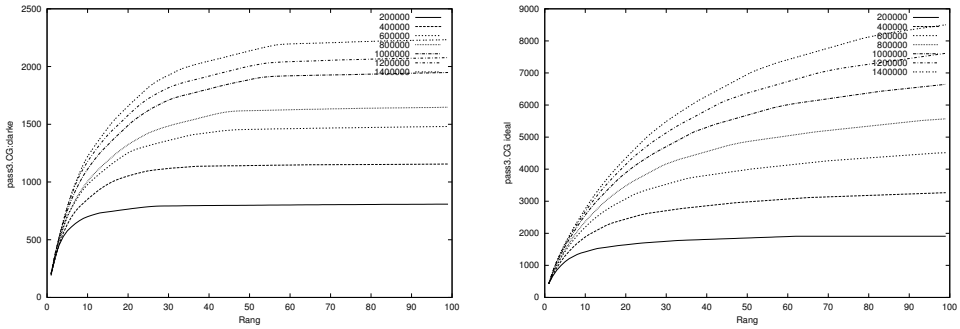


FIG. 5.29 – DCG avec 1/log2(rang), par taille de collection (modèle de *Clarke* et cas idéal) jusqu'au rang 100 Schéma (0, 5, 10).

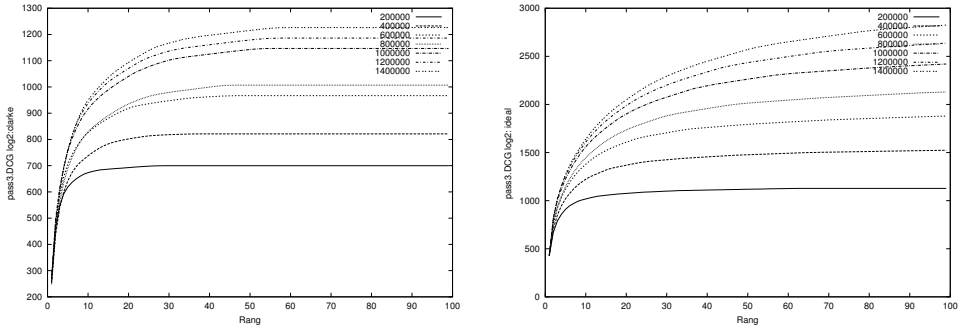


FIG. 5.30 – DCG avec 1/log210(rang), par taille de collection (modèle de *Clarke* et cas idéal) jusqu'au rang 100 Schéma (0, 5, 10).

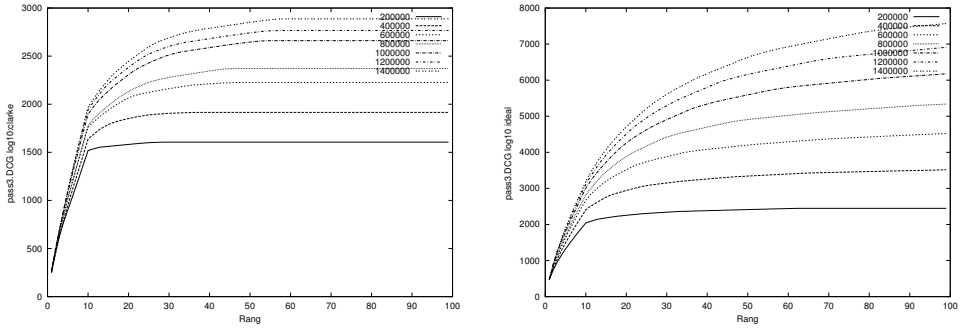


FIG. 5.31 – Métrique 1 sans cœfficient de pondération et avec 1/rang - cumul par rang - méthode de *Clarke* - Schéma (0, 5, 10).

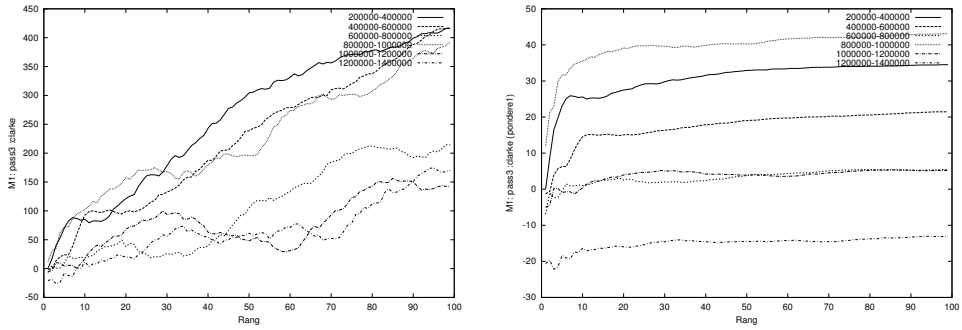


FIG. 5.32 – Métrique 1 avec les cœfficients de pondération 1/log2(rang) et 1/log10(rang) - cumul par rang - méthode de *Clarke* - Schéma (0, 5, 10).

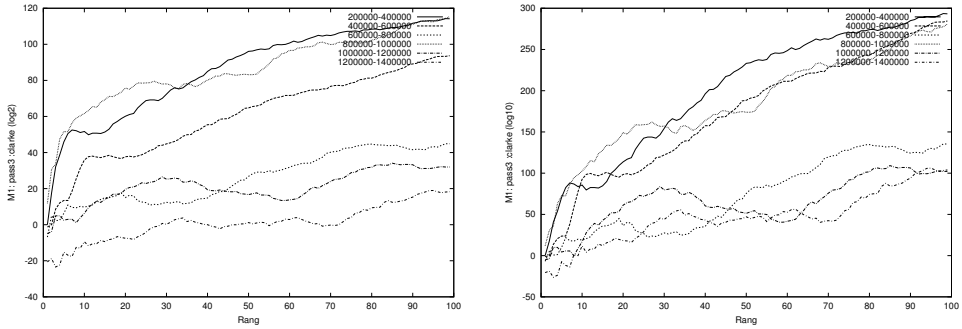


FIG. 5.33 – Métrique 2 sans cœfficient de pondération et avec 1/rang- cumul par rang - méthode de *Clarke* - Schéma (0, 5, 10).

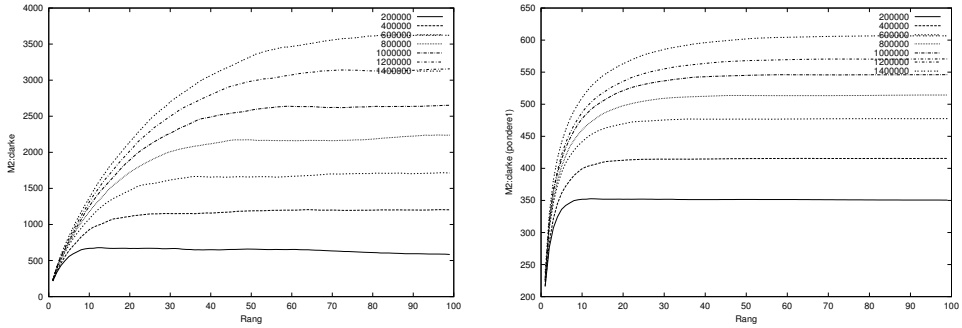
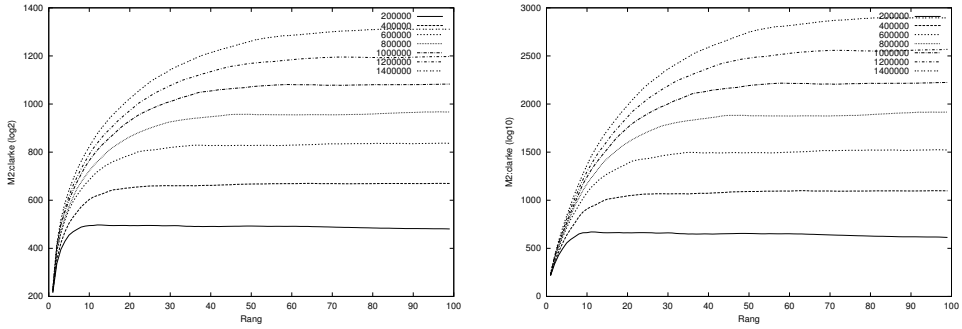


FIG. 5.34 – Métrique 2 avec les cœfficients de pondération 1/log2(rang) et 1/log10(rang) - cumul par rang - méthode de *Clarke* - Schéma (0, 5, 10).



croisent et chevauchent : on n'a pas de tendance ferme pour la métrique CG sur les tous premiers rangs. Sur des rangs plus grands, on note une augmentation du CG avec la taille de collection ; autrement dit le cumul des degrés de pertinence des documents retournés augmente avec la taille de la collection. La liste des résultats contient donc des documents avec des meilleurs niveaux de pertinence au fur et à mesure que la taille de collection augmente, si on ne se place pas en tout début de la liste des résultats. On note cette tendance sur tous les schémas de pondérations utilisés (0, 1, 2) et (0, 5, 10) et (0, 10, 100).

Le DCG fournit le cumul de gain d'information avec pondération par le rang au fil du parcours de la liste de résultats. Sur les tous premiers rangs (jusqu'à 5), on ne note pas une tendance particulière dans le comportement de la métrique DCG au fil de la croissance des collections sur les tous premiers rangs. Il y a des chevauchements entre les courbes correspondant aux différentes tailles de collections. Les différents cœfficients de pondération liés au rang utilisés changent l'évolution des courbes mais ne fournissent aucune tendance bien ferme au fil de la croissance des collections sur les tous premiers rangs. Ensuite la métrique DCG augmente globalement avec la taille de collection (les courbes des grandes collections sont presque confondues et se chevauchent légèrement).

Comme montré de façon théorique dans les sections précédentes, en utilisant certaines fonctions de distances pour calculer le gain d'information entre deux degrés de pertinence, nous obtenons l'égalité $Metrique1_N(C_1, C_2) = DCG_N(C_1) - DCG_N(C_2)$. Dans le cas d'utilisation de ces fonctions distances, la courbe de Métrique 1 est la « différence »entre la courbe de $DCG(C_1$ et la courbe de $DCG(C_1$. Les courbes des figures 5.8 et 5.9 illustrent cette égalité (pour le cas de la méthode de *Clarke*), valables évidemment pour toutes les méthodes. En effet, sur la figure 5.6, au rang 10, la différence entre la courbe de DCG pour la taille 400 000 et la courbe de DCG pour la taille 200 000 fournit la valeur de l'ordonnée obtenue sur la courbe $200\,000 - 400\,000$ de Métriques1 (figure 5.8) pour ce même rang 10.
Des valeurs négatives de Metriques 1 montrent une détérioration de résultats lors de la croissance en taille des collections, et des valeurs positives l'inverse. Sur les tous premiers rangs, on note que certaines courbes sont dans la zone négative dénotant les résultats décrits pour la métrique DCG. Par exemple, la courbe concernant les collections de 1 200 000 et 1 400 000 restent dans le négatif (on peut voir sur la figure 5.4 que la métrique DCG de la collection de taille 1 200 000 reste en effet légèrement au dessus de DCG de la collection de taille 1 200 000 documents). Les courbes des autres collections repassent dans le positif après les tous premiers rangs et y reste (sur les courbes de DCG y correspondant, on retrouve une amélioration globale de DCG avec la croissance des collections à partir de ces mêmes rangs).

L'utilisation des schémas de pondération accordant un poids très élevé au meilleur niveau de pertinence donne des résultats similaires pour la métrique CG et pour la métrique DCG. Pour cette dernière, on note toujours un chevauchement sur les tous premiers rangs ; pour les rangs suivants, l'ordre entre les courbes ne se fait pas strictement par taille de collection mais cet ordre reste stable sur tous les rangs (jusqu'à 100) ; les courbes des plus grandes collections sont proches entre elles et améliorent globalement de DCG lorsqu'on s'éloigne des tous premiers rangs.
Pour le cas du schéma de pondération (0, 10, 50, 100) qui ne correspond pas à des valeurs d'importance (et donc pour lequel il n'y a pas de courbes de CG et de DCG correspondantes), les Métriques 1 ont un comportement semblable à celui indiqué précédemment (voir figure 5.24 à figure 5.27). Ce qui montrent que dans ces cas aussi, le cumul des niveaux de pertinence des documents retournés pour les tous premiers rangs ne suit pas une tendance particulière avec la croissance en taille des collections ; sur les rangs suivants grands, les courbes des grandes collections sont proches les uns des autres et améliorent globalement ce cumul.
Ces métriques permettent donc d'interpréter effectivement des courbes liées à des schémas de gain d'information entre niveaux de pertinence que la seule attribution des valeurs d'importance ne permettrait pas de modéliser.

De même que pour les métriques Métriques 1, pour certains choix de fonctions de gain d'information entre deux degrés de pertinence, nous obtenons :
$Metrique2_N(C) = DCG_N(C) - DCG_N(C_{ideal})$. Les courbes 5.10, 5.11 illustrent cette égalité. Par exemple, en prenant la différence entre la valeur du DCG sur la courbe à 200 000 au rang 10 pour la méthode de *Clarke* (figure 5.6) et la valeur du DCG idéal au rang 10 pour 200 000 (même figure 5.6), nous obtenons la valeur de l'ordonnée de la courbe de Métriques2 au rang 10 pour 200 000 (figure 5.10).
Les Métriques 2 fournissent la différence entre les degrés de pertinence des documents de la liste des résultats et les degrés de pertinence de la liste idéale de résultats. Plus la différence tend vers 0, plus la liste de résultats est proche de la liste idéale. On remarque que les courbes pour les Métriques 2 se positionnent par ordre de taille des collections. Au fur et à mesure de la croissance des collections, les courbes s'éloignent de la droite $y = 0$ (droite qui traduirait une superposition avec la liste idéale et donc une différence nulle). Ainsi, les listes de résultats des petites collections sont plus proches de leurs listes idéales que les listes de résultats des grandes collections. Cette tendance se retrouve dans les courbes de Métriques 2 (si on se limite à une position fixe donnée). Aux positions fixes 5 et 10, le cumul des gains d'information (entre degrés de pertinence des documents de la liste des résultats et les documents de la liste idéale) augmente avec la taille de la collection, montrant ainsi une différence plus grande avec la liste idéale pour les grandes

collections (figure 5.12 à la figure 5.17).
Une explication réside dans la méthodologie de croissance des collections : elle impose une même proportion de documents par niveau de pertinence et par *topic* quelle que soit la taille de la collection. Ceci induit une augmentation du nombre de documents par niveau de pertinence et par *topic* au fil de la croissance des collections. Cette augmentation rend la liste idéale des résultats plus difficile à approcher au fil de la croissance des collections.

5.4 Synthèse

Dans ce travail, des métriques pour évaluer la manière dont les SRI passent à l'échelle sont proposées. Ces métriques s'appuient sur la notion de pertinence multivaluée. Leur but est de fournir des informations sur la cohérence entre l'ordre des documents retournés par un SRI et les degrés de pertinence de ces documents. Les métriques standards de RI sont basées sur une notion de pertinence binaire. Les métriques comme la *Discounted Cumulative Gain* ou la *Cumulative Gain* se basent également sur la pertinence multivaluée ; mais pour une collection donnée et un SRI, ces métriques calculent le gain d'information pertinente réalisée au fur et à mesure qu'on parcourt la liste des résultats retournés par ce SRI sur cette collection. Les métriques que nous proposons calculent plutôt le gain d'information pertinente réalisé lorsqu'un même SRI est utilisé sur différentes collections. Ainsi, en appliquant nos métriques sur des sous-collections de taille croissante d'une très grande collection, il est possible d'évaluer la capacité du SRI à classer les documents en fonctions de leur degré de pertinence lorsque l'on passe à l'échelle dans la taille de collection.
Le travail réalisé sur ce point de façon expérimentale avec différents SRI est également présenté. Les contraintes liées à l'attribution des valeurs numériques d'importance aux niveaux de pertinence ont été formalisées. La notion de gain d'information entre deux niveaux de pertinence est introduite et est utilisée pour définir des métriques concernant le passage à l'échelle. La mise en relation de nos métriques avec des métriques existantes prenant en compte plusieurs niveaux de pertinence a également été réalisée de façon théorique et expérimentale. Les notions de gain d'information entre niveaux de pertinence fournissent un cadre qui permet, en fonction des paramètres choisis, d'une part de retrouver les métriques existantes et d'autre part de prendre en compte des cas que ces métriques ne gèrent pas avec aise.

Chapitre 6

Conclusions et perspectives

Sommaire

6.1 Contexte de la thèse

La croissance exponentielle de l'information entraîne de nombreuses complications dans tous les domaines dont le cœur est la manipulation de l'information. Le domaine de la recherche d'information n'échappe pas aux effets de cette croissance. Cette thèse se situe dans ce dernier domaine et vise à étudier l'influence du passage à l'échelle sur les modèles de recherche d'information.

Dans cette thèse, l'accroissement de la quantité d'information numérique a été examiné au travers de ses causes et d'études quantitatives. Nous présentons les concepts de base de la recherche d'information notamment en parcourant les étapes classiques d'un processus de recherche d'information. Nous analysons l'impact du passage à l'échelle sur chacune de ces étapes en observant le comportement des moteurs de recherche (commerciaux) et au travers ds travaux antérieurs du domaine de la recherche d'information, ainsi que les différentes façons de le prendre en compte dans chacune de ces étapes. Ainsi, de la construction de l'espace de recherche à la phase de retour et visualisation des résultats de recherche, en passant par les étapes d'indexation et d'interrogation, le passage à l'échelle nécessite une réelle prise en compte afin de ne pas affecter de façon négative le processus de recherche. Cette prise en compte se fait de plus en plus dans les travaux de recherche. Les campagnes d'évaluation utilisent des collections de test de plus en plus

volumineuses, les techniques de visualisation des grandes quantités de données s'appliquent de plus en plus à la visualisation des résultats en recherche d'information, plusieurs travaux visent à rendre les phases d'indexation et d'interrogation plus rapides et plus efficaces. Deux phases ont retenu particulièrement notre attention : celle de mise en place de l'espace de recherche et celle d'évaluation.

La phase de mise en place de l'espace de recherche consiste en la collecte des informations au sein desquelles pourra s'effectuer une recherche. Pour étudier l'impact de passage à l'échelle, il semble opportun de disposer de divers espaces de recherche de taille croissante sur lesquels les performances des systèmes de RI seront analysées. Nos proposition porte sur une méthodologie pour la mise en place de tels espaces de recherche, permettant d'analyser ensuite l'effet du passage à l'échelle.

La phase d'évaluation des systèmes de recherche d'information permet de mesurer les performances des systèmes et éventuellement de les comparer. Cette phase constitue le cœur des différentes campagnes d'évaluation du domaine. Les performances des systèmes sont mesurées à l'aide de différentes métriques ayant toutes un concept en commun : le concept de la pertinence. Nous discutons de ce concept en dressant un état de l'art des travaux qui s'y sont intéressés suivant deux grandes approches : la pertinence comme concept binaire et la pertinence comme concept multivalué. Le premier est celui qui prévaut dans les campagnes d'évaluation et dans de nombreux travaux en RI. Toutefois, de plus en plus de travaux s'intéressent à la pertinence comme concept multivalué et cette dernière a de nombreux arguments qui la soutendent et que nous partageons. Nous avons donc proposé des métriques pour l'évaluation du passage à l'échelle utilisant la pertinence multivaluée.

6.2 Synthèse des contributions

En ce qui concerne notre premier axe de travail, notre proposition porte sur une méthodologie pour la construction d'espaces de recherche similaires et de taille croissante. La caractéristique de similarité des espaces de recherche construits dépend directement de ce que l'on souhaite observer quand on passe à l'échelle. Cette méthodologie se décline de deux manières :

- Une première technique permet de partir d'un espace de recherche volumineux, de le transformer en un espace de recherche dit uniforme par rapport à une caractéristique donnée. Ce second espace peut être découpé en sous-espaces de recherche de taille croissante, sans contrainte sur le découpage ; chacun des sous-espaces a les mêmes caractéristiques que l'espace de départ, ce qui permet d'analyser l'influence de la taille en s'affranchissant du biais

lié au contenu des sous-espaces de recherche.

- Une seconde technique qui est l'extension de la première consiste à construire des sous-espaces de recherche de taille croissante, en choisissant de façon aléatoire leur contenu, avec pour contrainte que la caractéristique d'intérêt soit la même sur chaque sous-espace construit. En construisant pour chaque taille un grand nombre de sous-espaces de recherche, le choix aléatoire de leurs contenus s'appuie sur l'utilisation des techniques statistiques connues comme la technique Monte-Carlo pour combiner les résultats et fournir des interprétations statistiquement fiables.

  Une mise en œuvre expérimentale de chacune de ces techniques a été réalisée. Notre cas d'étude est l'influence du passage à l'échelle sur les performances des systèmes de RI, performances calculées en utilisant les métriques classiques de RI (pertinence binaire), et la caractéristique de similarité est la proportion des documents pertinents. Ces expérimentations ont permis d'observer le comportement des métriques classiques utilisées en RI face au passage à l'échelle.

Dans le second axe de nos travaux, des métriques pour évaluer le passage à l'échelle dans des environnements à pertinence multivaluée ont été proposées. Pour ce faire, nous formalisons la notion de valeur d'importance attribuée à chaque degré de pertinence dans de tels environnements et nous introduisons la notion de gain d'information entre deux degrés de pertinence, qui permet de quantifier l'information pertinente que l'on gagne ou perd en parcourant des documents ayant ces degrés de pertinence. Ces deux notions permettent de proposer des métriques dont le but est de fournir des informations sur la cohérence entre l'ordre des documents retournés par un SRI et les degrés de pertinence de ces documents, quand le SRI travaille sur des espaces de recherche de taille croissante. L'utilisation de divers schémas pour le gain d'information entre deux degrés de pertinence agit directement sur le poids donné à un degré de pertinence par rapport aux autres. Il est ainsi possible de modéliser diverses configurations d'évaluation. Par exemple, une évaluation qui permet de reconnaître les systèmes de RI qui retournent peu de documents pertinents, mais retournent les documents de meilleur degré de pertinence en tête de leurs listes de résultats. Ou encore une évaluation permettant de reconnaître les systèmes qui retournent plutôt les documents de bon degré de pertinence. Le lien entre les métriques proposées et des métriques existantes comme la métrique *(Discounted) Cumulative Gain* qui utilisent également la pertinence multivaluée est mis en exergue. Pour certains schémas d'attribution des valeurs d'importance et d'attribution des gains d'informations entre degrés de pertinence, les métriques proposées peuvent être déduites directement de ces métriques existantes. Le gain d'information entre deux degrés de pertinence que nous introduisons permet toutefois de modéliser des cas de

figure qui ne peuvent pas être mis en oeuvre à travers ces métriques (par exemple le cas où le gain d'information entre deux degrés de pertinence successifs n'est pas le même à chaque fois, ou encore le cas de non « symétrie »dans la fonction de gain d'information).

6.3 Limites et Perspectives

Une des limites principales des travaux présentés dans cette thèse concerne la taille « réduite »des collections que nous avons à disposition pour nos expérimentations. Il serait intéressant de travailler sur des collections plus vastes comme la Terabyte. Ce type de collection n'a vu le jour que récemment dans les campagnes d'évaluation et pour l'instant, l'incomplétude des jugements de pertinence qui y sont liés est un problème non résolu.
De plus, la collection utilisée pour les expérimentations sur les métriques à pertinence multivaluée prend en compte deux degrés de pertinence pour un document et le nombre de documents par degré de pertinence est « faible »(notamment pour les documents de degré 2, en moyenne 4 par *topic* avec une médiane de 2). Une collection prenant en compte plusieurs degrés de pertinence avec un nombre plus grand de documents par *topic* pour chaque degré de pertinence permettrait de mieux observer l'effet des divers schémas d'attribution de valeur d'importance et de gain d'information entre degrés de pertinence (une telle collection a été utilisée dans les travaux de *Kekäläinen et Järvelin* mais ne sera mise à disposition qu'en 2007). Les collections utilisées dans des campagnes d'évaluation comme *NTCIR* ou *INEX* sont pour le premier cas de taille plus petite que la collection *WT10G* et pour le second cas délicates à utiliser dans notre cadre (documents structurés, construction de la liste idéale délicate (travaux de G. Kazai). Il serait toutefois intéressant de s'y pencher.

La méthodologie proposée pour la construction d'espaces de recherche similaires de taille croissante peut être appliquée avec des caractéristiques de similitude adaptées à d'autres cas d'étude, comme la répartition des termes de requêtes dans les documents.

Tout modèle de RI apparie une représentation des documents et une représentation des requêtes pour construire une liste de résultats de recherche. Il peut être intéressant de partir des formules d'appariement des deux représentations pour analyser de façon théorique ce que serait l'impact du passage à l'échelle (pour des modèles utilisant des paramètres liés à la taille de la collection par exemple ou les fréquences documentaires des termes). Pour ce faire, il faudra modéliser différentes sortes de croissance de collections (en terme de croissance du vocabulaire, fréquence des termes, taille des documents, etc.).

L'hypothèse de similarité de caractéristiques des collections au fil de l'augmentation en

nombre de documents nous a permis de nous affranchir du biais lié au contenu mais ne correspond pas forcément au mode de croissance réelle de la quantité d'information. Des croissances de collections avec de moins en moins de documents pertinents (travaux de *Voorhees* ayant abouti à la métrique *bpref*) ou alors avec une surabondance de documents pertinents dans les grandes collections permettraient d'observer le comportement des modèles de RI sous d'autres angles du passage à l'échelle.

Pour ce qui concerne la pertinence multivaluée, l'attribution d'une valeur d'importance à un niveau de pertinence reste un problème non résolu en RI. Il serait tout à fait intéressant de mener des travaux visant à proposer une classification des différents types de fonctions d'importance en fonction des applications.
Sur ce même volet, l'introduction d'un niveau de pertinence pour les documents non jugés dans les campagnes d'évaluation nous semble tout à fait judicieuse. Ces documents n'ayant en effet pas une probabilité certaine d'être non pertinents.
Dans nos travaux, pour un modèle donné, nous avons observé son comportement lors du passage à l'échelle. Une première continuation consisterait à mener des travaux permettant de comparer ce comportement face au passage à l'échelle en pertinence binaire et en pertinence multivaluée. Une seconde piste consisterait à comparer les comportements des différents modèles face au passage à l'échelle.

Nos travaux ont porté sur la phase de mise en place de l'espace de recherche et sur la phase d"évaluation. L'effet du passage à l'échelle sur les autres étapes du processus de recherche d'information mérite également d'être étudié de plus près.
L'influence et la prise en compte du passage à l'échelle est une problématique vaste qui dépasse largement le champ de la recherche d'information et touche de nombreux domaines manipulant les systèmes d'information.

Liste des notations

$gP = \sum_{d \in R} r(d)/n$	*Generalized non binary precision*
$gR = \sum_{d \in R} r(d) / \sum_{d \in D} r(d)$	*Generalized non binary recall*
$CG(i)$	*Cumulative Gain* au rang i
$DCG(i)$	*Discountad Cumulative Gain* au rang i
C_i	sous-collection numéro i
t	un *topic* t
$E(t)$	Ecart ou Distance théorique entre deux documents pertinents pour le *topic* t dans la collection uniforme
$Pert(t)$	ensemble des documents pertinents pour le *topic* t
$E_r(t)$	Ecart réel entre entre deux documents pertinents pour le *topic* t dans la collection uniforme
D	ensemble de tous les documents d'une collection
T	Ensemble de tous les *topics*
niv_i	niveau de pertinence i
$\succ$	relation d'ordre entre les niveaux de pertinence
$I(niv_i$	importance du niveau de pertinence niv_i
$g(niv_i, niv_j)$	gain d'information entre deux niveaux de pertinence
$d(x, y)$	distance mathématique
$NivPertinence(d_i)$	niveau de pertinence du document d_i
$d_k^t(C_i)$	document retourné au rang k lorsque la collection C_i est interrogée avec le *topic* t
$Passage_k^t(C_i, C_j)$	gain d'information entre les niveaux de pertinence des documents $d_k^t(C_i)$ et $d_k^t(C_j)$
$Retrieved_N^t(C_i)$	Liste des N premiers documents retournés lorsque la collection C_i est interrogée avec le *topic* t

$cp(k)$	cœfficient de pondération ; c'est une fonction décroissante du rang k
$Metrique1_N^t(C_i, C_j)$	cumul des gains d'information entre les niveaux de pertinence des documents de rangs égaux retournés après interrogation des collections C_i et C_j avec le *topic* t
$Documents^t(niv_i)$	ensemble des documents de niveau de pertinence niv_i pour le *topic* t
$Retrieved\_ideal(C)$	liste *idéale* des résultats qui devrait être retournée après interrogation de la collection C ; pour chaque *topic*, les documents sont classés par ordre décroissant suivant le niveau de pertinence pour *topic*

Table des figures

Glossaire

boxplot : boîte à moustaches ou boîte à pattes. Technique permettant de visualiser graphiquement une distribution de données en s'appuyant sur des statistiques d'ordre calculée sur celle-ci (médiane, quartiles par exemple).

Collection de test : Ensemble de documents, de requêtes et de jugements de pertinence faits sur les documents en utilisant ces requêtes.

Documents : Ce sont des entités qui encapsulent l'information. Elles sont minimales dans le sens où ce sont les résultats retournés par le système.

Echantillon(nage) : une partie d'un ensemble plus vaste qui en est représentative.

Efficience : la capacité de rendement, la performance, la rapidité.

Efficacité : le fait de produire l'effet attendu, est lié à la qualité des résultats.

Gain (d'information) : entre deux niveaux de pertinence, il fournit la qualité - quantité d'information qui différencie deux documents de ces deux niveaux de pertinence.

Haute précision : Métrique d'évaluation qui fournit la précision sur les premiers documents retournés.

Importance d'un niveau de pertinence : une valeur (numérique) attribuée à ce niveau de pertinence et qui permet d'établir une relation d'ordre entre les niveaux de pertinence.

Liste idéale : Liste de résultats au sein de laquelle tous les documents de la collection sont classés par niveau de pertinence décroissant vis à vis d'une requête spécifique.

Niveau de pertinence : le niveau de pertinence d'un document par rapport à une requête fournit le degré auquel ce document correspond à la requête. Dans le cadre de la pertinence binaire, les documents sont soit pertinents, soit non pertinents. Dans le cadre de la pertinence multivaluée, plusieurs niveaux de pertinence sont possibles pour un document.

Passage à l'échelle : la croissance ; dans notre cas la croissance des espaces de recherche.

Pertinence : la pertinence d'un document par rapport à une requête donne le degré de correspondance du document à la requête.

Précision : Métrique d'évaluation donnant la proportion de documents pertinents parmi ceux qui sont retournés.

Pooling : Technique utilisé dans les campagnes d'évaluation en RI pour construire les jugements de pertinence.

Rappel : Métrique d'évaluation donnant la proportion de documents retournés parmi ceux qui sont pertinents

Requête : Expression du besoin d'information de l'utilisateur (par exemple sous forme d'une liste de mots).

Score d'un document : valeur-système quantifiant l'appariement entre le document et une requête.

Topic : Besoin d'information ayant un numéro, un titre, un champ *description* qui fournit une description du sujet de ce besoin d'information, un champ *narrative* qui donne les caractéristiques attendues des documents attendus en réponse à ce besoin d'information.

Uniformisation : procédé par lequel nous construisons une collection sur laquelle une caractéristique donnée est la même quelle que soit la portion de la collection.

Index

Bibliographie

[1] http ://trec.nist.gov/.

[2] http ://www.technorati.com/.

[3] http ://inex.is.informatik.uni-duisburg.de :2003.

[4] http ://www-nlpir.nist.gov/projects/terabyte/.

[5] http ://www.seg.rmit.edu.au/lucy/.

[6] http ://www.seg.rmit.edu.au/zettair/.

[7] http ://www.cs.mu.oz.au/mg/.

[8] Ntcir workshop 1 : Proceedings of the first ntcir workshop on retrieval in japanese text retrieval and term recognition, tokyo, japan. In N. Kando and T. Nozue, editors, *NTCIR*, 1999.

[9] J. Aslam and M. Montague. Models for metasearch. In *Proceedings of the 24th ACM SIGIR conference on Research and Development in Information retrieval*, pages 276–284, 2001.

[10] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press Addison-Wesley, 1999.

[11] P. Bailey, N. Craswell, and D. Hawking. Engineering a multipurpose test collection for web retrieval experiments. *Information Processing and Management : an International Journal archive*, 39(6) :853–871, 2003.

[12] Y. Bar-Hillel. Summary of area 6 discussion. In *Proceedings of the International Conference on Scientific Information*, Washington DC, 1958.

[13] J. M. C. Bastien and D. Scapin. Ergonomic criteria for the evaluation of human-computer interfaces. Technical Report RT-0156, Institut national de recherche en informatique et en automatique, France, juin 1993.

[14] J. M. C. Bastien and D. L. Scapin. *Evaluation des systèmes d'informationet critères ergonomiques*, volume 2, pages 53–79. Paris, 2001.

[15] M. Beigbeder and A. Mercier. Etude des distributions de tf et de idf sur une collection de 5 millions de pages html. In *Atelier de recherche d'informations sur le passage à l'échelle Congrès INFORSID 2003*, Nancy, France, 2003.

[16] N. J. Belkin, C. Cool, W. C. Croft, and J. P. Callan. The effect of multiple query representations on information retrieval performance. In *Proceedings of the 16th Annual International ACM SIGIR conference on Research and Development in Information retrieval*, pages 339–346.

[17] N. J. Belkin, R. N. Oddy, and H. M. Brooks. Ask for information retrieval part i background and theory. *Journal of Documentation*, 38(2) :61–71, 1982.

[18] N. J. Belkin, R. N. Oddy, and H. M. Brooks. Ask for information retrieval part ii. results of a design study. *Journal of Documentation*, 38(3) :145–164, 1982.

[19] M. Bergman. The deep web : surfacing hidden value. *The Journal of Electronic Publishing*, 7(1), 2001.

[20] S.-A. Berrani. *Recherche approximative de plus proches voisins avec contrôle probabiliste de la précision : Application à la recherche d'images par le contenu.* PhD thesis, Université de Rennes 1, Rennes, France, fevrier 2004.

[21] M. Berry, Z. darmac, and E. Jessup. Matrices, vector spaces and information retrieval. *SIAM*, 41(2) :335–362, 1999.

[22] T. A. Brooks. The relevance aura of bibliographic records. *Information Processing and Management*, 33(1) :69–80, 1997.

[23] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2000.

[24] C. Buckley and E. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information retrieval*, pages 25–32, 2004.

[25] R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing and Management*, 28(5) :619–627, 1992.

[26] I. Campbell. Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assesments. *Journal Of Information Retrieval*, 2(1) :89–114, 2000.

[27] M. Chalmers and P. Chiston. Bead :explorations in information visualization. In *Proceedings the 15th ACM SIGIR conference on Research and Development in Information retrieval*, pages 330–337, 1992.

[28] J. P. Chevallet, J. Martinez, M. Boughanem, L. Lechani-Tamine, and S. Calabretto. Rapport final de l'as-91 du RTP-9 passage à l'échelle dans la taille des corpus, 2004.

[29] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe. Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems*, 20 :171–191, 2002.

[30] C. L. A. Clarke, G. Cormack, and E. Tudhope. Relevance ranking for one to three term queries. *Information Processing and Management*, 26(2) :291–311, 2000.

[31] C. L. A. Clarke and F. Scholer. The trec 2005 terabyte track. In *Proceedings of the 14th Text REtrieval Conference(TREC 2005)*, pages 166–174, 2005.

[32] C. W. Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. Technical report, 1970.

[33] C. W. Cleverdon, J. Mills, and E. M. Keen. *Factors determining the performance of indexing systems*, volume Two volumes. 1968. Two volumes.

[34] W. S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1) :19–37, 1971.

[35] G. Cormack, C. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 282–289, 1998.

[36] W. B. Croft. *Combining approaches to information retrieval.*

[37] C. A. Cuadra and R. V. Katter. Opening the black box of relevance. *Journal of Documentation*, 23(4) :291–303.

[38] C. A. Cuadra and R. V. Katter. The relevance of relevance assessment. In *Proceedings of the American Documentation Institute*, volume 4, pages 95–99, American Documentation Institute, Washington, DC, 1967.

[39] Cyveillance. Sizing the internet. http ://www.cyveillance.com/resources/library.asp, 2004.

[40] B. Dervin and M. S. Nilan. Information needs and users. *Annual Review of Information Science and Technology*, (21) :3–33, 1986.

[41] Dogpile. Different engines, different results. http ://comparesearchengines.dogpile.com/OverlapAnalysis.pdf, 2005.

[42] P. Dopichaj. The university of kaiserlautern at inex 2005. In L. N. in Computer Science, editor, *Advances in XML IR and evaluation : fourth workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005)*, 2005.

[43] S. Dumais. Lsi meets trec :a status report. In *Proceedings of the 1st Text REtrieval Conference*, pages 137–152, 1992.

[44] S. Dumartin and F. Mignard. L'informatique à la maison : une diffusion sensible, mais encore très ciblée. *INSEE Prémière*, (629), janvier 1999.

[45] C. E. and I. P. Dimensions of relevance. *Information Processing and Management*, 36 :533–550, 2000.

[46] R. A. Fairthorne. Summary of area 6 discussion. In *Proceedings of the International Conference on Scientific Information*, Washington DC, 1958.

[47] D. J. Foskett. A note on the concept of relevance. *Information Storage and Retrieval*, 8 :77–78, 1972.

[48] E. A. Fox. Characterization of two new experimental collections in computer science and information science containing textual and bibliographical concepts. Technical Report 83-561, http ://ncstrl.org, 1983.

[49] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval 7 Conference (TREC-2 , National Institute of Standards and Technology Special Publication 500-215*, pages 243–252, 1994.

[50] O. Frieder and D. Grossman. On scalable information retrieval systems. In *Invited Paper, 2nd IEEE International Symposium on Network Computing and Applications.*, Massachusett, Cambridge, 2003.

[51] T. J. Froehlich. Relevance reconsidered|towards an agenda for the 21stcentury : Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science*, 45(3) :124–133, 1994.

[52] Y. Frydel. Internet au quotidien : un français sur quatre. *INSEE Prémière*, (1076), Mai 2005.

[53] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3) :243–255, 1992.

[54] M. Fuller, E. Mackie, R. Sacks-Davis, and R. Wilkinson. structured answers for a large structured document collection. In *Proceedings of the 16th ACM SIGIR conference on Research and Development in Information retrieval*, pages 204–213, Pittsburgh, Pennsylvania, June 1993. ACM Press.

[55] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in

human-system communications. *Communications of the Association for Computing Machinery*, 30(11) :964–971, 1987.

[56] H. Gilbert and K. S. Jones. Statistical bases of relevance assessment for the "ideal" information retrieval test collection. Technical Report BL RD Report 5428, Computer Laboratory University of Cambridge, 1977.

[57] M. Gluck. Exploring the relationship between user satisfaction and relevance in information systems. *Information Processing and Management*, 32(1) :89–104, 1996.

[58] W. Goffman. On relevance as a measure. *Information Storage and Retrieval*, 2 :201–203, 1964.

[59] T. Grabs and H. J. Schek. Eth zürich at inex : Flexible information retrieval from xml with powerdb-xml. In *The First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*, Schloss Dagstuhl, Germany, December 2002.

[60] H. Greisdorf. Relevance : An interdisciplinary and information science perspective, 2000.

[61] C. Gurrin and A. Smeaton. Replicating web structure in small-scale test collections. *Information retrieval*, 7 :239–263, 2004.

[62] H. T. H. and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3) :187–222, 1991.

[63] S. P. Harter. Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9) :602–615, 1992.

[64] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1) :37–49, 1996.

[65] S. P. Harter and C. A. Hert. Evaluation of information retrieval systems : approaches, issues, and methods. *M. E. Williams(ed.) Annual Review of Information Science and Technology*, 32 :3–94, 1997.

[66] D. Hawking and N. Craswell. Very large scale retrieval and web search. In E. Voorhees and D. Harman, editors, *TREC : Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[67] D. Hawking and S. Robertson. On collection size and retrieval effectiveness. *Information retrieval*, 6(1) :99–105, 2003.

[68] D. Hawking and P. Thistlewaite. Proximity operators - so near and yet so far. In *Proceedings of the Fourth Text Retrieval Conference TREC-4*, pages 131–143, 1995.

[69] D. Hawking and P. Thistlewaite. Scaling up the trec collection. *Information retrieval*, 1(1) :115–137, 1999.

[70] D. Hawking, E. Voorhees, N. Craswell, and P. Bailey. Overview of the trec8 web track. In *Proceedings of the 8th Text REtrieval Conference*, pages 131–150, 1999.

[71] M. A. Hearst and C. Karadi. Cat-a-one : an interactivinterface for specifying searches and viewing retrieval results using a large category hierarchy. In *Proceedings the 20th ACM SIGIR conference on Research and Development in Information retrieval*, pages 246–255, 1997.

[72] S. Heinz and J. Zobel. Efficient single pass index construction for text databases. *Journal of American Science on Information and Technology*, 54(8) :713–729, 2003.

[73] R. Hendley, N. Drew, A. Wood, and R. Beale. Narcissus : visualising information. In *Proceedings of Information Visualisation Symposium*, pages 90–96, 1995. N. Gershon and S. G. Eick IEEE CS Press.

[74] W. Hersh. Relevance and retrieval evaluation : perspectives from medecine. *Journal of the American Society for Information Science*, 45(3) :201–206, 1972.

[75] W. R. Hersh, C. Buckley, T. J. Leone, and D. H. Hickman. Ohsumed : An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th International ACM SIGIR conference on Research and Development in Information retrieval*, pages 192–201, 1994.

[76] D. J. Hillman. The notion of relevance. *American Documentation*, pages 26–34, January 1964.

[77] D. L. Howard. Pertinence as reflected in personal constructs. *Journal of the American Society for Information Science*, 45(3) :172–185, 1994.

[78] A. Imafouo and M. Beigbeder. Scalability influence on retrieval models : An experimental methodology. In *27th European Conference on Information Retrieval*, pages 388–402, 2005.

[79] P. Ingwersen. *Information retrieval interaction*. Taylor Graham Publishing, London, UK, 1992.

[80] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23th annual international ACM SIGIR conference on Research and Development in Information retrieval*, pages 41–48, 2000.

[81] K. S. Jones. *Information retrieval experiment*. Butterworth-Heinemann, 1981.

[82] K. S. Jones and R. Bates. Report on a design study for the "ideal" information retrieval test collection. Technical Report BL RD Report 5481, Computer Laboratory University of Cambridge, 1979.

[83] K. S. Jones and K. V. Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32 :59–75, 1976.

[84] Y. Kagolovsk and J. Moehr. Current status of the evaluation in information retrieval. *Journal of medical systems*, 27(5) :409–424, October 2003.

[85] J. Katzer, M. J. McGill, J. Tessier, W. Frakes, and P. Dasgupta. A study of the overlap among document representations. *Information Technology :Research and Development*, 1(2) :261–274, 1982.

[86] G. Kazai and M. Lalmas. Notes on what to measure in inex. In *INEX Workshop on Element Retrieval Methodology*, 2005.

[87] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53(13) :1120–1129, 2002.

[88] J. Kekäläinen. Binary and graded relevance in ir evaluations -comparison of the effects on rankings of ir systems. *Information Processing an Management*, 41, 2005.

[89] C. C. Kuhlthau. *Seeking meaning : A process approach to library and information science*. Ablex Publishing, Norwood, NJ, 1993.

[90] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi. Pooling for a large-scale test collection : An analysis of the search results from the first ntcir workshop. *Information Retrieval*, 5(1) :41–59, 2002.

[91] F. W. Lancaster and A. J. Warner. *Information Retrieval Today*. Information Resources Press, Arlington, VA, USA, 1993.

[92] S. Lawrence and C. L. Giles. Searching the world wide web. *Science*, 280 :98–100, 1998.

[93] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature Magazine*, 400(6740), 1999.

[94] J. H. Lee. Combining multiple evidence from different properties of weighting schemes. In *Proceedings of the 18th Annual International ACM SIGIR conference on Research and Development in Information retrieval*, pages 180–188, 1995.

[95] J. H. Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th International ACM SIGIR conference on Research and Development in Information retrieval*, pages 267–276, 1997.

[96] M. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4 :343–359, 1969.

[97] H. P. Luhn. A statistical approach to mechanical encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4) :309–317, 1957.

[98] P. Lyman, H. R. Varian, K. Swearingen, P. Charles, N. Good, L. L. Jordan, and J. Pal. How much informations 2003. http ://www.sims.berkeley.edu/research/projects/how-much-info-2003/, October 2003.

[99] D. M. Mackay. What makes the question. *The Listener*, 63(5) :789–790, 1960.

[100] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *SIGIR 2001 Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–275, 2001.

[101] R. Manmatha and H. Sever. A formal approach to score normalization for metasearch. In *Human Language Technology Conference HLT 2002*, 2002.

[102] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7 :216–244, 1960.

[103] Y. Mass and M. Mandelbrod. Retrieving the most relevant xml components. In *Proceedings of the second Workshop of tne INitiative for the Evaluation of XML retrieval (INEX)*, pages 53–58, Schloss Dagstuhl, Germany, December 2004.

[104] M. McGill, M. Koll, and T. Norreault. An evaluation of factors affecting document ranking by information retrieval systems. Technical report, Syracuse, Syracuse University School of Information Studies, 1979.

[105] C. T. Meadow. *Text Information Retrieval Systems.*

[106] A. Mercier. Etude comparative de trois approches utilisant la proximité entre les termes de la requête pour le calcul des scores des documents. In *INFORSID 2004 - 22ème congrès informatique des organisations et des systèmes d'information et de décision*, pages 95–106, 2004.

[107] A. Mercier and M. Beigbeder. Calcul de la pertinence basée sur la proximité pour la recherche d'informations. *Document Numérique*, 9(1) :43–60, 2006.

[108] S. Miyamoto. *Fuzzy Sets in information retrieval and cluster analysis.* 1990.

[109] S. Mizzaro. How many relevances in information retrieval ? *Interacting with Computers*, 10(3) :303–320, 1998.

[110] A. Moffat and J. Zobel. Self-indexing inverted files for fast text retrieval. *ACM Transactions on Information Systems*, 14(4) :349–379, 1996.

[111] C. Mooers. Zatocoding applied to mechanical organization of knowledge. *American Documentation*, 2 :20–32, 1951.

[112] G. B. Newby. The science of large scale information retrieval. Internet archives, 2000.

[113] L. Nigay and F. Vernier. Design method of interaction techniques for large information spaces. In *Proceedings of AVI-98*, pages 37–46, 1998.

[114] M. S. Nilan, R. P. Peek, and H. W. Snyder. A methodology for tapping user evaluation behaviors : An exploration of users' strategy, source and information evaluating. In C. L. Borgman, Pai, and E. Y. H., editors, *Proceedings of the American Society for Information Science (ASIS) 51st Annual Meeting*, pages 152–159, Atlanta, G. A. Medford, NJ, October 1988.

[115] J. OConnor. Relevance disagreements and unclear request forms. *American Documentation*, pages 165–177, July 1967.

[116] K. A. Olsen, R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams. Visualisation of a document collection : the vibe system. *Information Processing and management*, 29(1) :69–81, 1993.

[117] M. L. Pao. Term and citation retrieval : A field study. *Information Processing and Management*, 29(1) :95–112, 1993.

[118] C. Peters, editor. *Cross Language Information Retrieval and Evaluation, workshop of cross-language evaluation forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised papers*, volume 2069 of *Lecture Notes in Computer Science*. Springer, 2001.

[119] D. Raitt. Digital libraries initiatives across europe. *Computers in libraries*, 20(10) :26–35, 2000.

[120] Y. Rasolofo and J. Savoy. Term proximity scoring for keyword-based retrieval systems. In *Proceedings of European Conference on Information Retrieval Research*, pages 207–218, 2003.

[121] A. M. Rees. The relevance of relevance to the testing and evaluation of document retrieval systems. 1966.

[122] A. M. Rees and D. G. Schulz. A field experimental approach to the study of relevance assessments in relation to document searching. 2 vols. Technical Report NSF Contract No. C-423, Center for Documentation and Communication Research, School of Library Science, 1967.

[123] R. R.Korfhage. *Information Storage and Retrieval*. Wiley, 1997.

[124] S. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33 :294–304, 1977.

[125] T. Sakai. Average gain ratio : A simple retrieval performance measure for evaluation with multiple relevance levels. In *Proceedings of the ACM SIGIR conference on Research and Development in Information retrieval*, 2003.

[126] G. Salton. *The SMART retrieval system : experiments in automatic document processing*. Automatic Computation. Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[127] G. Salton. A theory of indexing. In *Regional Conference series in Applied Mathematics*, 1975.

[128] G. Salton, A. A. Fox, and H. Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11) :1022–1036, 1983.

[129] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1983.

[130] M. Sanderson and H. Joho. Forming test collections with no system pooling. In *SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40, New York, NY, USA, 2004. ACM Press.

[131] B. Sandore. Online searching : What measures satisfaction ? *Library and Information Science Research*, 12(1) :33–54, 1990.

[132] T. Saracevic. Relevance : A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26 :321–343, 1975.

[133] T. Saracevic. Evaluation of evaluation in information retrieval. In *Proceedings of the 18th ACM SIGIR Conference on research and development in information retrieval*, pages 138–146, July 1995.

[134] T. Saracevic. Relevance reconsidered. In P. Ingwersen and N. O. Pors, editors, *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2). Information science : integration in perspective*, pages 201–218, Copenhagen (Denmark), october 1996.

[135] T. Saracevic. Evaluation of digital libraries :an overview. In *Workshop on the Evaluation of Digital Libraries*, 2004.

[136] T. Saracevic and P. Kantor. A study of information seeking and retrieving. iii. searchers, searches, overlap. *Journal of the American Society for Information Science*, 39(3) :197–216, 1968.

[137] T. Saracevic, P. Kantor, A. Y. Chamis, and D. Trivison. A study of information seeking and retrieving. 1. background and methodology. *Journal of the American Society for Information Science*, 39(3) :161–176, 1988.

[138] L. Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29 :3–48, 1994.

[139] L. Schamber, M. B. Eisenberg, and M. S. Nilan. A re-examination of relevance : toward a dynamic, situational definition. *Information Processing and Management*, 26(6) :755–776, 1990.

[140] T. Schlieder and H. Meuss. Querying and ranking xml documents. *Journal of the American Society for Information Science and Technology*, 53(6) :489–503, April 2992.

[141] F. Scholer, H. Williams, J. Yiannis, and J. Zobel. Compression of inverted indexes for fast query evaluation. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information retrieval*, pages 11–15, 2002.

[142] H. Small. The relationship of information science to the social sciences : a cocitation analysis. *Information Processing and Management*, 17(1) :39–50, 1981.

[143] S. Smithson. The evaluation of information retrieval systems : A case study approach. In ed. K. F. Jones, editor, *Informetrics 10 Prospects for Intelligent Retrieval, Proceedings of Aslib Conference*, pages 75–89, London, 1990.

[144] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM Conference on research and Development in Information Retrieval*, pages 66–73, 2001.

[145] E. Sormunen. *A method for measuring wide range performance of Boolean queries in full-text databases*. PhD thesis, University of Tampere, Tampere, Finland, 2000.

[146] E. Sormunen. Liberal relevance at trec - counting on negligible documents ? In *Proceedings of the 25th Annual International ACM SIGIR conference on Research and Development in Information retrieval*, 2002.

[147] A. Spink, H. Greisdorf, and J. Bateman. From highly relevant to not relevant : examining different regions of relevance. *Information Processing and Management : an International Journal*, 34(5) :599–621, 1998.

[148] L. T. Su. Evaluation measures for interactive information retrieval. *Information Processing and Management*, 28(4) :503–516, 1992.

[149] L. T. Su. The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*, 45(3) :207–217, 1994.

[150] J. M. Tague-Sutcliffe. Some perspectives on the evaluation of information retrieval systems. *Journal of the American Society of Information Science*, 47(1) :1–3, 1996.

[151] R. Tang, W. M. Shaw, and J. L. Vevea. Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science*, 50(3) :254–264, March 1999.

[152] M. Taube. A note on the pseudomathematics of relevance. *American Documentation*, 16(2) :69–72, 1965.

[153] R. S. Taylor. Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29 :178–194, 1968.

[154] M. A. Tiamyu and I. Y. Ajiferuke. A total relevance and document interaction effects model for the evaluation of information retrieval processes. *Information Processing and Management*, 24(4) :391–404, 1988.

[155] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information and Processing Management*, (36) :697–716, 2000.

[156] E. Voorhees and D. Harman. Overview of the eight text retrieval conference. In *Proceedings of the 8th Text REtrieval Conference*, pages 1–24, 1999.

[157] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information retrieval*, pages 74–82, 2001.

[158] D. W and G. Marchionini. A comparative study of web search service performance. In *Proceedings of The Annual Conference of the American Society for Information Science*, pages 136–142, 1996.

[159] P. Wallis and J. A. Thom. Relevance judgements for assessing recall. *Information Processing and Management*, 32(3) :273–286, 1996.

[160] P. Wang and M. D. White. A cognitive model of document use during a research project. study ii. decisions at the reading and citing stages. *Journal of the American Society for Information Science*, 50(2) :98–114, 1999.

[161] P. Wilson. Situational relevance. *Information Storage and Retrieval*, 9(8) :457–471, 1973.

[162] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes - Compressing and indexing documents and images*. Morgan Kaufman Publishers, second edition, 1999. http ://www.cs.mu.oz.au/mg/.

[163] S. Wu and F. Crestani. Methods for ranking information retrieval systems without relevance judgements. In *SAC 2003*, 2003.

[164] G. K. Zipf. *Human Behavior and the principle of least effort*. Addison-Wesley Press, Cambridge, MA, 1949.

[165] J. Zobel. How reliable are the results of large scale information retrieval experiments. In *Proceedings of the 21th ACM SIGIR conference on Research and Development in Information retrieval*, pages 307–314, 1998.

Résumé

Les évolutions technologiques de ces dernières années ont entraîné une croissance exponentielle de la quantité d'information numérique disponible. La Recherche d'Information, discipline dont le cœur de metier est la manipulation de cette information est questionnée par cette croissance rapide. Les travaux présentés dans cette thèse se sont penchés sur le problème de l'influence du passage à l'échelle sur les performances des modèles de Recherche d'Information. Après un tour des travaux du domaine qui prennent en compte le passage à l'échelle, une méthodologie pour construire des espaces de recherche de tailles croissantes et dont le contenu est contrôlé est proposée dans un premier temps ; ces espaces de recherches sont utilisés pour observer les performances de divers modèles de RI en fonction de la taille des données manipulées. Dans un second temps, les travaux portent sur la proposition de métriques prenant en compte plusieurs niveaux de pertinence pour un document ; la notion d'importance d'un niveau de pertinence est formalisée et la notion de gain d'information entre deux niveaux de pertinence est introduite. Ces deux notions permettent de fournir des métriques dédiées à analyser la capacité des systèmes de RI à retourner des documents en fonction de leur niveau de pertinence, au fur et à mesure que la taille de l'espace de recherche augmente.

Mots-clés: Processus et modèles de Recherche d'Information, passage à l'échelle, collections et sous-collections, évaluation, métriques, pertinence (multivaluée)

Abstract

Information grows continuously ; for professional or personal reasons the need of easy access to it comes under the Information Retrieval field. We first of all make a tour of IR works linked to the scalability, and we notice that few of theses works tackled the questions of Information Retrieval Systems *effectiveness* in the context of scalability in corpus size. After that, the first part of our work is about a methodology which makes it possible to study the scalability influence on some properties of IR models. This methodology constructs a succession of collections of growing sizes on which a given characteristic C (that acts on the studied properties) is the same ; then we analyze the properties as the collection size increases.

The second part of our work relates to metrics for evaluating the ability of IRS to rank documents according to their relevance levels when collection size increases. Indeed, for the

user's point of view, in large environments, it can be desirable to have Information Retrieval Systems that retrieve documents according to their relevance levels. Relevance levels have been studied in some previous Information Retrieval works while some others (few) IR research works tackled the questions of IRS effectiveness and collections size. These latter works used standard IR measures on collections of increasing size to analyze IRS effectiveness scalability. In this part of our work, we bring together these two issues in IR (multigraded relevance and scalability) by designing these metrics.

Keywords: Information retrieval process and models, scalability, collections and sub-collections, evaluation, metrics, (multigraded) relevance

Références de l'auteur

Revue nationale avec comité de sélection

A.IMAFOUO et M. BEIGBEDER Vers des protocoles d'évaluation du passage à l'échelle. *Revue Ingéniérie des Systèmes d'Information (ISI).* Vol 11, Numéro 4/2006

Conférences internationales avec comité de sélection

A.IMAFOUO and M. BEIGBEDER Evaluating scalability in Information Retrieval with multigraded relevance. *Proceedings of the Asian Information Retrieval Symposium(AIRS).* October 2006. SINGAPORE

A.IMAFOUO and M. BEIGBEDER An experimental methodology to study collections size impact on retrieval effectiveness. *Proceedings of the Dutch Belgian Workshop on Information Retrieval (DIR'05)*, January 2005. Utrecht, HOLLAND.

A.IMAFOUO and M. BEIGBEDER Scalability influence on retrieval effectiveness :An experimental methodology. *Proceedings of the European Conference on Information Retrieval (ECIR'05).* March 2005. Santiago De Compostela, SPAIN.

A.IMAFOUO et X. TANNIER Retrieval Status Value in IR evaluation. *Proceedings of the String Processing and Information Retrieval (SPIRE 2005)*, November 2005. Buenos Aires, ARGENTINA.

Conférences nationales avec comité de sélection

A.IMAFOUO et M. BEIGBEDER Evaluer le passage à l'échelle dans des environnements à pertinence multivaluée. *Actes de la 3ème Conférence Francophone en Recherche d'Information (CORIA'06)*, Mars 2006. Lyon, FRANCE.

A.IMAFOUO et M. BEIGBEDER Passage à l'échelle : Une méthodologie d'étude de l'influence du volume de collection sur les modèles de Recherche d'Information. *Actes de la 2ème Conférence Francophone en Recherche d'Information (CORIA'05)*, Mars 2005. Grenoble, FRANCE.

A. MERCIER, A.IMAFOUO et M. BEIGBEDER Modèle de proximité : Conception

et comparaison à une méthode de recherche de passages. *Actes de la 2ème Conférence Francophone en Recherche d'Information (CORIA'05)*, Mars 2005. Grenoble, FRANCE.

Campagne d'évaluation

A. MERCIER, A.IMAFOUO et M. BEIGBEDER ENSM-SE at CLEF 2005 : Uses of Fuzzy Proximity Matching Function. *Working Notes for the CLEF 2005 Workshop*, September 2005. Vienna, Austria.

Poster et Ateliers internationaux avec mentor

A.IMAFOUO A scalability survey in Information Retrieval and Digital Libraries. *European Conference on Digital Libraries (ECDL 2005)*, September 2005. Vienna, AUSTRIA.

A.IMAFOUO Influence du passage à l'échelle sur modèles de Recherche d'Information. *INFORSID'04 (Forum Jeunes Chercheurs)*, Mai 2004. Biarritz, FRANCE.

Ecole Nationale Supérieure des Mines
de Saint-Etienne

N° d'ordre: 421 I

Name : AMELIE IMAFOUO

Thesis title : Studying the way Information Retrieval models scale

Speciality : Computer Science (Information Retrieval)

Keywords : Information retrieval process and models, scalability, collections and sub-collections, evaluation, metrics, (multigraded) relevance

Abstract

Information grows continuously; for professional or personal reasons the need of easy access to it comes under the Information Retrieval field. We first of all make a tour of IR works linked to the scalability, and we notice that few of theses works tackled the questions of Information Retrieval Systems effectiveness in the context of scalability in corpus size.

After that, the first part of our work is about a methodology which makes it possible to study the scalability influence on some properties of IR models. This methodology constructs a succession of collections of growing sizes on which a given characteristic C (that acts on the studied properties) is the same; then we analyze the properties as the collection size increases.

The second part of our work relates to metrics for evaluating the ability of IRS to rank documents according to their relevance levels when collection size increases. Indeed, for the user's point of view, in large environments, it can be desirable to have Information Retrieval Systems that retrieve documents according to their relevance levels. Relevance levels have been studied in some previous Information Retrieval works while some others (few) IR research works tackled the questions of IRS effectiveness and collections size. These latter works used standard IR measures on collections of increasing size to analyze IRS effectiveness scalability. In this part of our work, we bring together these two issues in IR (multigraded relevance and scalability) by designing these metrics.

Ecole Nationale Supérieure des Mines
de Saint-Etienne

N° d'ordre : 421 I

Prénom Nom : Amélie IMAFOUO

Titre de la thèse : Etude de l'influence du passage à l'échelle sur les modèles de recherche d'information

Spécialité : Informatique

Mots clefs : Processus et modèles de Recherche d'Information, passage à l'échelle, collections et sous-collections, évaluation, métriques, pertinence binaire et multivaluée.

Résumé

Les évolutions technologiques de ces dernières années ont entraîné une croissance exponentielle de la quantité d'information numérique disponible. La Recherche d'Information, discipline dont le cœur de métier est la manipulation de cette information est questionnée par cette croissance rapide. Les travaux présentés dans cette thèse se sont penchés sur le problème de l'influence du passage à l'échelle sur les performances des modèles de Recherche d'Information.

Après un tour des travaux du domaine qui prennent en compte le passage à l'échelle, des méthodologies pour construire des espaces de recherche de tailles croissantes et dont le contenu est contrôlé sont proposées dans un premier temps; ces espaces de recherches sont utilisés pour observer les performances de divers modèles de RI en fonction de la taille des données manipulées.

Dans un second temps, les travaux portent sur la proposition de métriques prenant en compte plusieurs niveaux de pertinence pour un document; la notion d'importance d'un niveau de pertinence est formalisée et la notion de gain d'information entre deux niveaux de pertinence est introduite. Ces deux notions permettent de fournir des métriques dédiées à analyser la capacité des systèmes de RI à retourner des documents en fonction de leur niveau de pertinence, au fur et à mesure que la taille de l'espace de recherche augmente.